DOCUMENT RESUME

ED 188 587 IR 008 440

AUTHOR McGill, Michael
TITLE An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems.
INSTITUTION Syracuse Univ., N.Y. School of Information Studies.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Oct 79
GRANT NSF-IST-78-10454
NOTE 144p.

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Algorithms; *Comparative Analysis; Data Bases; Data Processing; Information Needs; *Information Retrieval; Search Strategies; *User Satisfaction (Information)
IDENTIFIERS *Boolean Algebra; *Document Ranking

ABSTRACT

This study of ranking algorithms used in a Boolean environment is based on an evaluation of factors affecting document ranking by information retrieval systems. The algorithms were decomposed into term weighting schemes and similarity measures, representatively selected from those known to exist in information retrieval environments, before being tested on documents submitted by specific clearinghouses to the CIJE database. Searches were conducted using information need statements from individuals with interests congruent to the database, and documents retrieved by professional searchers were judged for relevance by those submitting the need statements. This input was analyzed according to the ability of the algorithms to move relevant documents toward the beginning of the output list using the coefficient of ranking effectiveness (CRE). While it is possible to significantly improve the order of the output using either a controlled vocabulary or free text, the ranking is at best about 20 percent effective at this time. It is suggested that the factors currently used in ranking algorithms are not likely to make ranking closer to 100 percent effective. A bibliography of 58 references is included. (Author/RAA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED188587

AN EVALUATION OF FACTORS AFFECTING DOCUMENT RANKING BY INFORMATION RETRIEVAL SYSTEMS

October 1979

Principal Investigator
Michael McGill

Research Associates
Matthew Koll
Terry Noreault

This material is based upon research supported by the National Science Foundation, Division of Information Science and Technology under Grant NSF-IST-78-10454. The opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation.

School of Information Studies
Syracuse University
Syracuse, New York 13210

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Michael J. McGill

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR008440

ABSTRACT

This is a report of a study of ranking algorithms used in a Boolean environment. The ranking algorithms are decomposed into term weighting schemes and similarity measures. Representative term weights and similarity measures are selected from those known to exist in information retrieval environments. The ranking algorithms are tested using documents submitted by specific clearinghouses to the Current Index to Journals in Education data base.

The study used information need statements from individuals with interests congruent with the data base. After searches were conducted by professional searchers, the retrieved documents were judged for relevance by the persons submitting the original information need statement. This provided the input to study the ranking algorithms.

The algorithms were analyzed according to their ability to move relevant documents toward the beginning of the output list. The coefficient of ranking effectiveness (CRE) was used to measure this ability. The study found that when using a controlled vocabulary or the free text, it is possible to significantly improve the order of the output. The results also indicate that ranking is at best about 20% effective with the remaining 80% not yet resolved. It is suggested that the factors currently used in ranking algorithms are not likely to make ranking closer to 100% effective. Rather, new information is likely to be required.

ACKNOWLEDGEMENTS

This report constitutes the final report for Grant NSF-IST-78-10454, entitled AN EVALUATION OF FACTORS AFFECTING DOCUMENT RANKING BY INFORMATION RETRIEVAL SYSTEMS. The project was supported by the Division of Information Science and Technology of the National Science Foundation. The grant period ran from September 1, 1978 to February 29, 1980.

The study required the resources of many individuals including the individuals providing information need statements and relevance judgements. Peggy Montgomery was a prime coordinator, counselor, and typist for the project. Chris Fox's simulation study is a valuable contribution. The analysis was made possible through the guidance of Jeffrey Katzer, with assistance from Rich Veith and Chris Fox. The final report was edited by Cheryl McAfee. The able advice of Jennifer Kuehn, Gerard Salton and Karen Sparck Jones also significantly contributed to this study.

TABLE OF CONTENTS

			Page
I	-	Introduction	1
II	-	Components and Environment of Ranking Algorithms	3
		a. Form of Document Representation (DR)	3
		b. Term Weighting in the Document Representation.	6
		c. Form of the Query (QF)	7
		d. Term Weighting in the Query (TW)	8
		e. Similarity Measure (SM)	8
III	-	Review of Relevant Literature	10
IV	-	Approach and Methodology	19
V	-	Restrictions	21
VI	-	Methodological Requirements	22
VII	-	Objectives of the Study	23
VIII	-	Procedure	24
IX	-	Review and Selection of TWs	25
X	-	Review and Selection of SMs	34
XI	-	Description and Loading of the Data Base	61
		a. Introduction	61
		b. Description of Data Base	61
		c. Construction of the Inverted Files	63
		d. Comparison of Construction Times	65
XII	-	Collecting Interest Statements	68
XIII	-	Intermediaries and Searching	69
XIV	-	Query Processing	72
		a. Lost Responses	78
XV	-	Measuring the Effectiveness of Ranking	79
XVI	-	Overview of Results	85

TABLE OF CONTENTS, continued

			Page
XVII	-	Results Using the Controlled Representation	86
XVIII	-	Results Using the Free Representation	92
XIX	-	Efficiency Considerations	96
XX	-	Cost of TWs	98
		a. Category 1	99
		b. Category 2	100
		c. Category 3	100
XXI	-	Cost of SMs	100
XXII	-	Conclusion	102
		References	106
		Appendix A - Commands for SIRE	111
		Appendix B - Forms	114

TABLES

		Page
Table 1 -	Term Weightings	28
Table 2 -	Similarity Measures	36
Table 3 -	Unique SMs After Binary Simulation	58
Table 4 -	Records in Data Base By Clearinghouse	62
Table 5 -	Description of Data Base by Representation	63
Table 6 -	Processing Time for File Construction	66
Table 7 -	Comparison of File Sizes	66
Table 8 -	Summary of Search Characteristics	74
Table 9 -	Overlap Percentages	76
Table 10 -	Matrix of $\overline{CRE}$ Values Controlled Representation	87
Table 11 -	Analysis of Variance Results Term Weighting Scheme Controlled Representation	88
Table 12 -	Analysis of Variance Results Similarity Measure Controlled Representation	89
Table 13 -	Significant Differences between $\overline{CRE}$ Means for Similarity Measures Controlled Representation	90
Table 14 -	Matrix of $\overline{CRE}$ Values Free Representation	93
Table 15 -	Analysis of Variance Results Term Weighting Schemes Free Representation	94
Table 16 -	Analysis of Variance Results Similarity Measures Free Representation	95
Table 17 -	Significant Differences Between $\overline{CRE}$ Means for Similarity Measures Free Representation	97

FIGURES

		Page
Figure 1 -	Ranking Algorithm Model	4
Figure 2 -	Inverted File	6
Figure 3 -	Performance Measure Difficulty	13
Figure 4 -	Correlations Between SMs Based on Binary Simulation of 15 Queries	56
Figure 5 -	Output from READIN	67
Figure 6 -	Output from MERGE/SORT	67
Figure 7 -	Output from MAKDIC	67
Figure 8 -	Calculation of the Coefficient of Ranking Effectiveness	84

INTRODUCTION

Ranking the output of an information system on the criterion of probable relevance to the query is considered an optimal strategy in information retrieval (Maron and Kuhns, 1960; Gebhardt, 1975; Lancaster and Fayen, 1973). Ranked output attempts to provide the user with information indicating that the closer a document is to the beginning of the output list, the more likely it is to be relevant to his query.

In the context of a document retrieval system [1] a ranking algorithm defines an ordering on a set of documents, ordering the documents according to their degree of similarity to a query. In the simplest case (a binary decision rule) the document set is ordered into two classes; those satisfying the retrieval criteria and those not satisfying the criteria. More complicated ranking algorithms may be defined in order to provide orderings of greater detail, creating up to N classes of output, where N is the number of documents retrieved.

The absence of a systematic collection, classification, and comparison of ranking algorithms was noted by Sager and Lockeman (Sager and Lockeman, 1976). Subsequently they began the task of systematically exploring ranking algorithms. They

---

1 The phrase "document retrieval" will be used although "computerized reference retrieval system" is a more appropriate description of this type of system.

identified components which could theoretically be combined to form 990 different ranking algorithms; unfortunately only 14 of these algorithms could actually be tested given their experimental conditions. Their results were constrained by the fact that the ranking algorithms were testable in only one retrieval environment (defined below) and they encountered other difficulties, such as problems with relevance judgments (Sager and Lockeman, 1976, p. 24). Thus, work on the careful examination of ranking algorithms was begun, but much theoretical and empirical work remained.

Ranked output has been a process of unknown effectiveness, requiring heavy user effort or occurring only in the context of SMART-like systems. In the SMART-like systems, the ranking process cannot be isolated from the retrieval process for a particular investigation. Inverted file systems, on the other hand, traditionally keep information which only allows simple ranking and often requires a great deal of user effort.

Innovations in the Syracuse Information Retrieval System (SIRE) (McGill, et. al., 1976), allow numerous and sophisticated ranking methods to be studied in an inverted file context. Specific components of ranking algorithms and the retrieval environment can be isolated and simultaneously varied so that the efforts of each component and each combination of components can be observed.

This is a report of a study which examined 504 different

ranking algorithms in two different environments. The study was conducted from September 1, 1978 through August 31, 1979. The report considers the environment and the components of ranking algorithms, the relevant literature, the methods used in this study and the results and implications of the collected data.

COMPONENTS AND ENVIRONMENT OF RANKING ALGORITHMS

A fundamental model of a document retrieval system is shown in Figure 1. This model is not new, but it provides an essential framework for the understanding of ranking algorithms. There are many other models which view the information retrieval process from other perspectives with different components (e.g. Saracevic, 1968, p. xii). The model in Figure 1 is different in that the components and place of ranking algorithms in a retrieval system are clearly included. This model clarifies the relationships between certain processes. For example, it shows the complementary and analogous roles of defining a query for an information need and the role of defining the set of descriptors for a document. From this model it is clear that the schemes for applying different weights to terms in the query or document representations can be isolated from each other and from the formation of the representations. Further, the model clarifies the isolation of the similarity measure from the weighting and descriptions. The five key elements comprising a ranking algorithm and its environment are:

a) FORM OF DOCUMENT REPRESENTATION (DR). Form of document

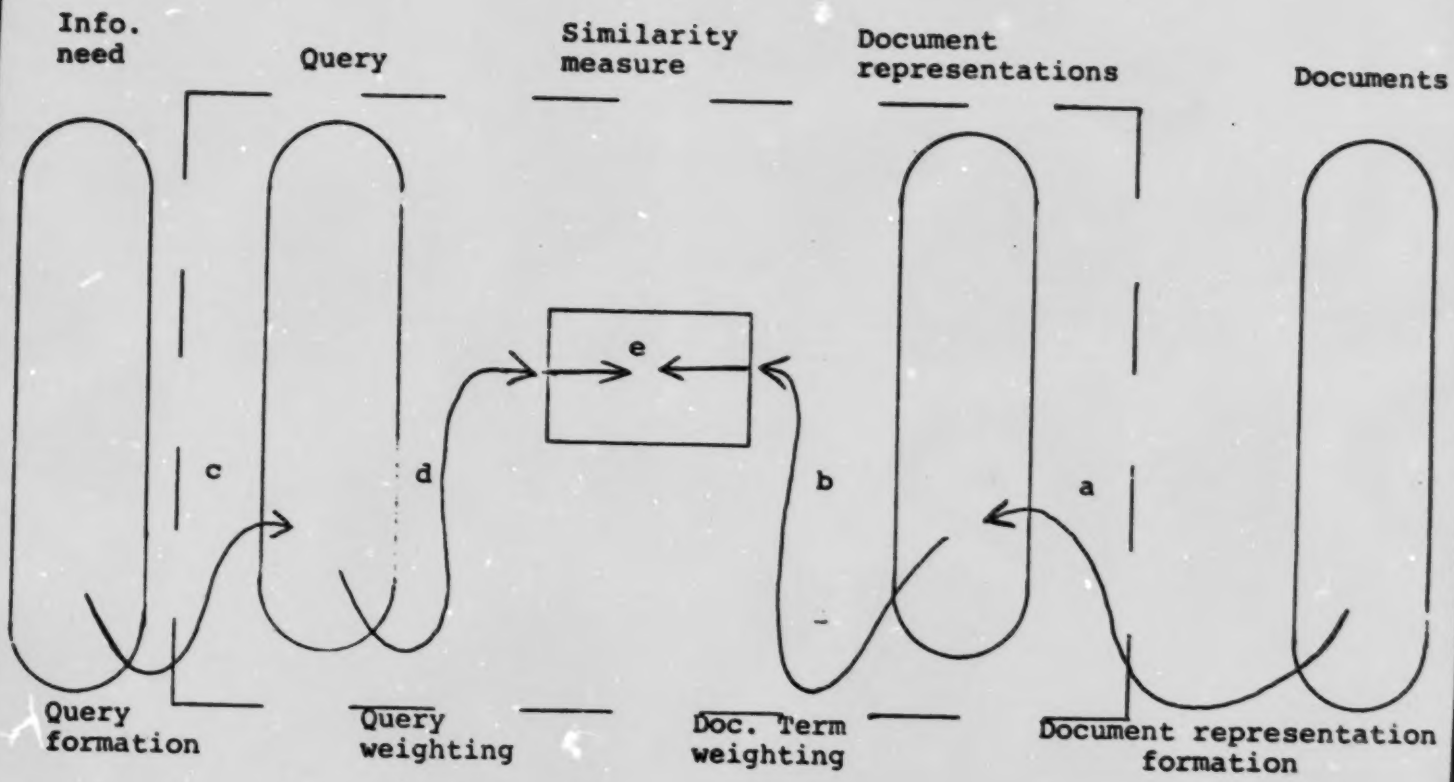


FIGURE 1

RANKING ALGORITHM MODEL

representation refers to the manner of selecting the descriptors (index terms) by which the document will be examined by an algorithm to determine whether to retrieve or not to retrieve the document. Indexing language variables constitute a large portion of Document Representation variability.

A document can be represented by any combination of classification codes or index terms assigned manually or automatically from a controlled or uncontrolled vocabulary. Terms may be extracted from the document or portions of the document (e.g. title, author, abstract, citations). There are also variations in the form, structure, depth and breadth of indexing languages.

The product of the document representation process, for any document representation and any document, j, can be thought of as a document vector $D_j = a_{1j}\ a_{2j}\ a_{3j}\ \ldots a_{mj}$, where

$$a = \begin{cases} 0 & \text{if document j is not indexed by term i} \\ 1 & \text{if document j is indexed by term i.} \end{cases}$$

m= number of index terms in the vocabulary

$1 \leq i \leq m$

Term	Doc. 1	Doc. 2	Doc. 3	.....Doc. N
Apex	1	0	0	 0
Apple	0	1	1	 0
Baker	1	0	1	 1
Bun	0	0	1	 1
.	.	.	.	..... .
.	.	.	.	..... .
.	.	.	.	..... .
.	.	.	.	..... .
.	.	.	.	..... .
term M	0	1	0	 1

FIGURE 2 - INVERTED FILE

The vector terminology (e.g. "document vector") is applicable to inverted file systems as well as to SMART systems. In Figure 2 a row is referred to as a "term vector" - noting the presence or absence of a particular term across all documents. A column is a "document vector" - noting the presence or absence of each term in a particular document. The process of indexing a document may thus be described as the creation of a document vector representing that particular information item.

b) TERM WEIGHTING IN THE DOCUMENT REPRESENTATION (TW). Term weighting schemes determine how much emphasis is placed on the occurrence(s) of each index term. Sager and Lockeman identified 22 such schemes (Sager and Lockeman, 1976). The elementary weighting scheme is, of course, "unweighted". This

scheme assigns a 1 or 0 for the presence or absence of a term, respectively. More complex schemes may count the number of occurrences of the term in the document, normalized by such factors as the number of terms used to represent the document, the frequency with which a term occurs throughout the data base, the overall frequency distribution and probabilities of the term occurrences, or the term's pattern of co-occurrence with other terms. The term weights may be described by a vector of coefficients $(w_{1j}\ w_{2j}\ w_{3j}\ \ldots w_{mj})$ for the corresponding document representation vector. A weighted document representation vector $D_j = w_1 a_{1j}\ w_2 a_{2j}\ w_3 a_{3j}\ \ldots\ w_m a_{mj}$ is developed by the element by element multiplication of the two vectors. Additionally, the elements of the document representation vector do not necessarily have to represent index terms themselves, but could be underlying factors discovered by analysis of the text of the collection (see Switzer, 1965) or other selected attributes (see Cleveland, 1976).

c) FORM OF THE QUERY (QF). Analogous to the conversion of a document to a document representation, an information need must be converted into a query. Queries are categorized here as belonging to one of two forms, Boolean or "natural" language[2]. Naturally, the query formation process results in a request expressed in the same language as the document representation.

---

2 Natural language queries may be considered as identical to a Boolean query consisting of the same set of terms, all connected by ORs with some subsequent processing (McGill et al. 1976).

Thus, there are theoretically two query forms (Boolean and natural language) for each document representation form. In either case the query can be represented by a vector corresponding to the document representation vector (i.e. having the column vectors represent the same terms, factors or attributes). Other factors which may influence the query formation include the use/non-use of an intermediary, the form of the man-machine dialog, relevance feedback techniques, thesauri, adjacency operators and generality/specificity of the query.

d) TERM WEIGHTING IN THE QUERY (TW). Weighting coefficients may be assigned to query vector elements as they are to document representation vectors. Query terms can be weighted equally, according to their frequency of occurrence in the query, manually, according to the searcher's perceptions of the importance of each term, or, in situations using relevance feedback, as a function of the term's pattern of occurrences in relevant and non-relevant documents (Yu and Salton, 1977).

e) SIMILARITY MEASURE (SM). A similarity measure is an algorithm which computes the degree of agreement between entities. For this study, the concern is with a query vector and document representation vectors. There are many vector association measures described in the literature, (see for example, van Rijsbergen, 1975, 31-34; Reitsma, 1968). One simple measure yields a 1 if $\sum_{i=1}^{m} (Q_i.D_i) \neq 0$ and 0 if $\sum_{i=1}^{m} (Q_i.D_i) = 0$ where Q and D are the query and document representation vectors and m is the number of terms in the vocabulary. More complex measures may take into account terms not present in either the query or

document representation vector in addition to number of terms, frequency and probability data.

Ranking algorithms are composed of the three units, in Figure 1, QW, TW and SM. The two units DR and QF constitute the environment of a ranking algorithm. The object of a ranking algorithm is to predict the relevance of each document and place the documents in descending order to predicted relevance. Thus, as the output list is read from beginning to end, each document is more likely to be relevant than those following it.

Ranking algorithms do not alter the composition of the set of documents retrieved by a query. That is, in a given environment (QF and DR) and a given data base, a document retrieval system will produce the identical set of documents in response to a query regardless of the ranking algorithm employed, provided that a cutoff value on the similarity score is not being used to restrict the size of the output list. Conceptually, ranking algorithms work after the retrieved set is formed to effect the order in which the documents are displayed.

This is precisely the way the two-step retrieval process has been implemented in the Syracuse Information Retrieval Experiment (SIRE) (McGill, et al., 1976; Noreault et al., 1977). First the retrieved set is identified as those documents which satisfy the Boolean logic of the query. Then the ranking algorithm is employed to compare the similarity of the document representation vectors of the retrieved documents to

the query vector. The documents are then rank ordered for output. In a recent study, the efficiency and effectiveness of this method were demonstrated (Noreault et al., 1977).

This process is in contrast to linear associative processing retrieval systems (e.g. unclustered SMART) (Sparck Jones, 1973). In these systems, as in the case of two-step systems using a cutoff value, the nature of the ranking algorithms can affect the set of documents the user received. Using a cutoff value places greater importance on the role of the ranking algorithm.

REVIEW OF RELEVANT LITERATURE

While there have been numerous evaluations of document retrieval systems and different aspects of document retrieval systems, Sager and Lockeman's (1976) view that a systematic evaluation (or even conceptual organization) of ranking algorithms has been absent from the literature is confirmed.

There are methodological reasons why definite statements about the relative performances of ranking algorithms were not made. These will be discussed below. However, a significant reason for the lack of knowledge is theoretical. That is, until Sager and Lockeman's explication, the concept of ranking algorithms was not well enough defined to be carefully examined.

It is easy to be critical of the methodology in information retrieval research. Without going through a case by case exam-

ination of past evaluation studies some recurrent problems can be pointed out. First is the problem of the small size of data bases usually used in this research. This inhibits the generalizability of results because the queries used in these studies are often not representative of queries that would be made of a larger data base. For example, consider a data base of 1,000 documents, and a query which retrieves 30 of those documents. If the data base is a representative sample of a larger data base, with say 30,000 documents, then that same query passed against the larger data base should be expected to retrieve about 900 documents - not the kind of query often used in operational settings. Another problem is that of human variables confounding system variables. Examples are poor indexer reliability, searcher inconsistency, and poor agreement between and among user and expert relevance judges.

For example, SUPARS researchers concluded that a large portion of the variance in system performance may be due to factors extrinsic to the system - the manner in which documents are defined as relevant and individual differences among searchers (Katzer, 1971, pp. 38-39). The Comparative Systems Laboratory group concluded that

> ... The difference in retrieval as exists between languages of equivalent length can almost entirely be attributed to human decisions in indexing and question analysis. In the study of differences in retrieval of relevant answers, where the relevant answers retrieved by index file C and missed by file F were examined, it was found that at least 75% of the incidence of missing can be attributed to the human factor - to human decisions, idiosyncrasies, inconsistencies, interpretations, etc. (Saracevic, 1968, p. 130).

Keen (1973) reported 42% inter-indexer consistency, 77% intra-indexer consistency (after 20 weeks), 32% agreement among relevance judges and that 69% of the documents judged relevant by a requestor were also judged relevant by expert judges.

Other methodological reasons for the lack of success of evaluations to explain ranking algorithms are 1) an examination of ranking algorithms has not been the main goal of any empirical research besides Sager and Lockeman's restricted effort, 2) in some cases the system variables have not been isolaced or controlled so as to determine if there are effects due to specific system components. This may be due both to the experimental design and/or the nature of the dependent variables (performance measures), and 3) that variables contributing to ranking algorithm performance have not been considered in broad contexts. That is, these variables must be considered at different levels of component and environmental variables.

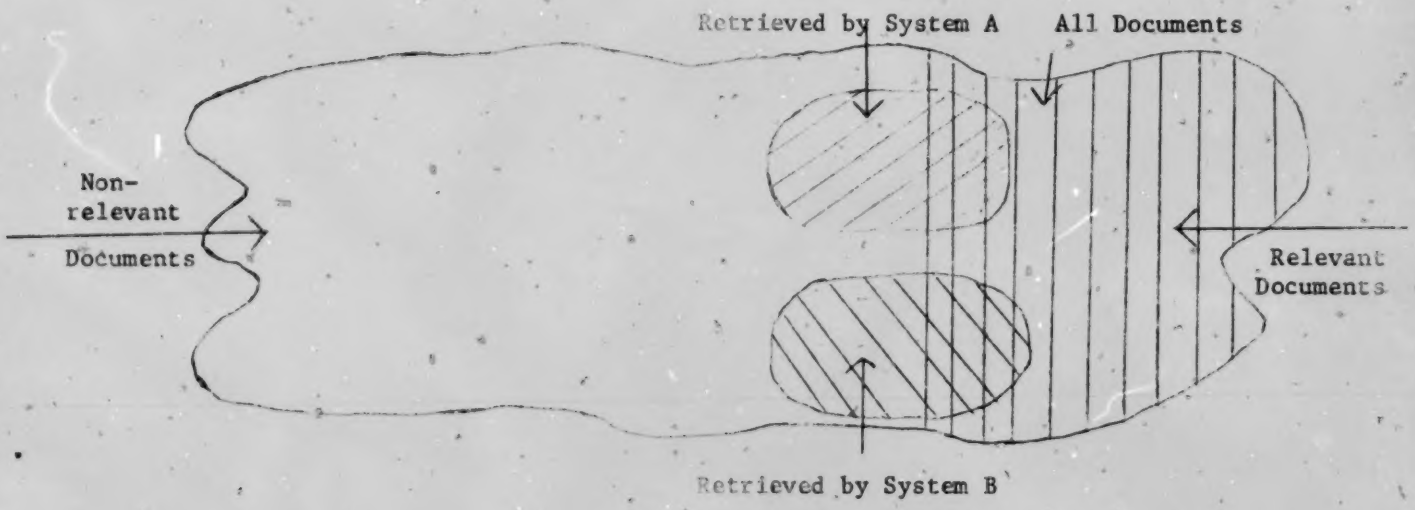
Evaluations or comparisons of total systems are too general to allow conclusions to be drawn about specific system components. This is particularly true for studies of operational systems, but true for experimental systems also. For example, in the original SMART vs. MEDLARS comparison (Salton, 1969) while manual and automatic indexing were the focus of the comparison, other factors such as the form of the query, term weighting schemes and similarity measures may have had some effect on the results.

Recall, precision and fallout measures may be suitable as descriptive measures of a system's overall performance, but they

are inadequate for investigations of the effects of specific systems components. These measures are sensitive to variance in many system components; thus it is only with the greatest experimental control that these measures can give testimony to the performance of a system component.

For example, consider Figure 3. Retrieval methods A and B could have identical recall - precision graphs yet be retrieving entirely different sets of documents. Usual performance measures would not convey that information, which is of value to a system designer or evaluator.

FIGURE 3 - PERFORMANCE MEASURE DIFFICULTY



The most prevalent problem is that of restricted range of investigation. Forms of document representation, index term weighting, similarity measures and query modifications have all been scrutinized. Unfortunately, in most instances, the other variables are ignored or insufficiently considered (cf. Reitsma

and Sagalyn, 1968; Minker et al., 1972; Cleverdon and Keen, 1966; Keen, 1973; Sparck Jones, 1973; Salton, 1975). This does not imply that such studies have not achieved results which bear on the performance of ranking algorithms. Studies of index term weighting (e.g. Salton and Yang, 1973; reviewed by Sparck Jones, 1973) show ambiguous results but indicate that inverse document frequency and discrimination value are valuable weighting factors, and that term frequency and document length may be useful variables.

Studies of document representation have found significant performance differences due to index language variables (Saracevic, 1968; Cleverdon and Keen, 1966; Keen, 1973). A sample of results from document representation studies indicate that uncontrolled vocabularies work as well as controlled vocabularies, that single term languages are superior to other types, that there may be an optimal depth for indexing languages and that machines and humans are generally better at judging relevance when they are given more text to work with (e.g. title vs. full text). In general, studies have found indexing languages to be a variable of minor importance (see Saracevic, 1968, pp.119-130).

Yet Swets looked at 50 different retrieval methods over three different systems (four different collections) and found that there were very small differences in the performance of different retrieval methods within a collection as opposed to the larger performance differences between collections. These differences are attributed to the difference in "hardness" of the vocabularies in the subject area of the collections and to differences

in the ways relevance judgements were made (Swets, 1967, p.28).

Studies of similarity measures have generally concluded that there are only minor differences among their performances. The cosine correlation has become the preferred measure (Reitsma and Sagalyn, 1968; van Rijsbergen, 1975). Still, conclusive tests of similarity measures for performance differences have not been conducted. Similarity measures have been studied as measures of association in the context of clustering items in a vector space rather than as a query-document matching function.

One must regard all of these results with caution. Due to the restricted ranges of other variables within which the key variables were tested, it is unknown if observed differences would remain consistant in different environments and if apparently equivalent methods behave differently in non-similar settings.[3] In other words, there might be interactions among variables complicating one's ability to discern key effects.

One is skeptical of "no difference" findings. While the variables may not have a difference on the employed performance measures, one might detect an effect on other dependent variables. (See example on Page 4, Figure 1.) There is likely to be a great deal of noise present in experiments. Given the

---

3 Saracevic's (1968) study is somewhat of an exception - a step in the direction of the currently proposed work. In his study, the variables "source of input" and "form of index language" were covaried. However, different term weighting schemes and similarity measures were not used.

factors that influence recall-like measures, plus human variance in indexing and relevance assessments, it would not be surprising to find significant differences overwhelmed by noise.

It should be noted that a simulated document ranking and cutoff procedure was used in the Cranfield II Study (Cleverdon and Keen, 1966) to determine if that procedure would affect the performance of index languages that were being studied. Unfortunately, the study was not a study of ranking algorithms. It was executed by hand, using only one ranking method, and was based on search co-ordination levels rather than textual statistical data. The data base consisted of 200 documents. A recall-based performance measure was used which was not suited to comparing ranking algorithms.

Sager and Lockeman (1976) defined the ranking algorithm composed of a query term weighting scheme, a document term weighting scheme and a similarity measure.[4] This conceptual structure, as mentioned previously, is vital to the study of the ranking process. They identified some 22 term weighting functions for documents, 5 query term weighting schemes and 9 similarity measures, yet this list was not exhaustive. Also, the algebraic relationships among ranking algorithm components required further exploration. Lerman found that many similarity measures are monotonic with respect to each other. (Lerman, 1970, cited in van Rijsbergen, 1975, p. 31.)

---

4 They also included the fourth trival phase of placing the documents in descending order of similarity to the query.

Sager and Lockeman (1976, p. 17) note that "Ranking algorithms cannot improve the results of retrieval, but only those of display." They assume a two-step model of retrieval (without cutoff) as described earlier. Systems in which the set of documents given to the user consists of all documents having a non-zero relationship with a natural language query meet the above definition. However, some ranking algorithms define an ordering such that all documents have some relation to the query (e.g. distance in a multidimensional space. See Katter, 1967; Switzer, 1965). In fact, the only situation in which ranking algorithms exist functionally independent of the retrieval set formation is when the set is formed by the search logic, and ranking occurs afterward. Thus Sager and Lockeman focused on the process of ranking the output from Boolean queries (this is not meant to exclude other logical operators).

In contrast to the previously mentioned support shown for ranked output, it has been argued that there are logical fallacies in the ranking of output from Boolean searches (Bookstein and Cooper, 1976; Bookstein, 1977) and that ranking options have not been utilized by users on systems which had them available (McCarn, 1976; Rickman, 1972).

The second point will be dealt with first. The ranking methods which went unused required considerable effort on the part of the user to manually assign weights or priorities to query terms or to make other related judgments. It should be noted that the methods explored in this study require no extra user

effort beyond that which would be part of a conventional Boolean search. Also, until the SIRE ranked output study (Noreault et al., 1977) there was little evidence that the output from Boolean searching could be effectively ranked according to probable relevance.

As for logical perils, Bookstein notes that any known system of ranking Boolean search output has logical inconsistencies due to the fact that the same query could be represented by different Boolean statements which would result in the same ranking method producing different document orderings in response to the same query (conceptually the same query). Also, inconsistency may arise from the fact that it is unclear how to deal with ANDS, ORS, and NOTS; specifically, does satisfying different requirements of the logic mandate different weights? What about the absence of a word when it is NOT supposed to be present? How much should that count?

These criticisms of ranking algorithms are logically correct. However, the documents are not being ranked according to their degree of agreement with the search statement. Documents either do or do not meet the requirements for inclusion in the retrieved set. The documents are then ordered along a useful dimension - in the case of the 1977 SIRE study, by degree of similarity to a vector composed of the terms used in the query. Any criteria that seek to measure the degree of relation to some aspect of the information need, or in any way predict relevance are valid to explore and use. Bookstein correctly asserts that

to rank documents on degree of conformity to the logic of a Boolean query is logically inconsistent. In fact, to do so without a probabilistic or fuzzy logic designed for that purpose would be logically incoherent.

Second, any ranking method is employed not as part of a scientific theory of meaning or logic, but as a pragmatic tool to aid in the satisfaction of an information need. Logical consistency is not required of many human tasks; satisfactory performance is required.

The Noreault et al. (1977) study referred to is an example of an empirical examination of a ranking algorithm other than Sager's and Lockeman's work. Noreault et al. found that a completely automatic algorithm was able to rank the output from Boolean searches effectively on probable relevance with no extra user effort and little incremental system cost. In that study the environment of the ranking algorithm was characterized by Boolean queries created by an intermediary and a stemmed free text vocabulary from titles and abstracts with 150 common words removed. The ranking algorithm consisted of query terms weighted by their number of occurrences in the query, document terms weighted by their frequency of occurrence in the title and abstract, and the cosine correlation as the similarity measure.

APPROACH AND METHODOLOGY

The approach taken here embodies a philosophy towards the

study of document retrieval systems. The study of information retrieval is in its infancy. There are fundamental aspects of computerized document retrieval systems about which little is known. Studies of overall system performance and user satisfaction are, of course, valuable. But similar emphasis needs to be placed on the functioning of various system components.

An emphasis on isolating and testing specific system components does not dictate studying each individual component in an isolated environment. One of the important aspects of this research design was the plan to isolate, control and vary the levels of several system component variables at the same time so that main effects and interactions could be studied.

Sager and Lockeman's three component model was expanded for this study to include two environmental classes of variables (Figure 1). Just as it was important to vary the levels at which ranking algorithm component variables combine so that a statement could be made about relative ranking algorithm performances within the particular environment in which they were tested, it was important to test the ranking algorithm combinations in various environments.

For example, in an environment in which all document representations are the same length, it makes little sense to employ a term weighting scheme or similarity measure that normalizes by the length of the document representation. Frequency with which a term occurs in a document representation is likewise not a

meaningful variable in a vocabulary of subject index terms or classification codes assigned by indexers.

Unfortunately, there are too many variables within the five variable classes in the model and too many levels of all of these variables to encompass in a single comprehensive test. Further, this is a study of the ranking process, not the entire retrieval process. So, some environmental factors have been simplified for the study.

RESTRICTIONS

The query form used in the study was Boolean queries. There are several reasons for this. 1) The study was designed to impact system designers working with the current state-of-the art. The vast majority of operational systems today provide for Boolean searching. Thus, in terms of query form, our results should be generalizable to that population. 2) The study measured the effect of ranking algorithms on already formed sets. As mentioned previously, the natural language systems used ranking algorithms to define these sets. 3) Methodologically, the measurement of the effectiveness of competing ranking algorithms becomes difficult if natural language queries are included. Natural language queries may retrieve sets of documents so large that a cutoff must be used to restrict the number of documents the user has to examine. This poses problems for comparison of retrieved sets. Also, no query weighting schemes which required user assigned weights were used.

The study was performed on a commercially available data base, Current Index to Journals in Education. Document representation forms were selected from those existing on the data base.

There is always a question about the generalization of results obtained from experimentation on a single data base. Cooper (1970) warns information retrieval researchers about the excessive number of variables to be considered. Swets' (1967) observation that the performance differences between collections was far greater than those between different retrieval methods within a collection was noted earlier. One expects that there will be no dramatic changes in different collections in the ranking algorithms found most effective. The effects are likely to be attributable to factors reflected in the document representation variable. However, replications in different collections would lend credence to the stability of the results.

METHODOLOGICAL REQUIREMENTS

The nature of the dependent variables (performance measures) used to test the effects of the ranking algorithms is an important consideration. Any measures used must specifically measure the ranking algorithms' effect and not reflect other aspects of the system. The measure should test only the change in ordering due to the ranking algorithms. For ease of understanding, it is also desirable that the measure be a single number rather than a curve. The Coefficient of Ranking Effectiveness (CRE) as described in Noreault (1977) was designed for this purpose.

An essential factor in this study was Syracuse Information Retrieval Experiment (SIRE). Its augmented inverted file design (see McGill et al., 1976) allowed the two-step processing of queries using a variety of ranking algorithms (QWs, TWs, and SMs). STAIRS has a comparable capability but is less flexible in this regard. Sager and Lockeman used STAIRS and were unable to vary QW or SM or use seven of the 22 TWs they described (Sager and Lockeman, 1976, p. 18).

OBJECTIVES OF THE STUDY

1) To assess the relative effectiveness of alternative methods of ranking the output from Boolean queries, (that is, alternative methods of predicting the relative relevance of retrieved documents). Specifically,
    a) To assess the effectiveness of various term weighting schemes (within and across DRs);
    b) To assess the effectiveness of various similarity measures (within and across DRs).
2) To determine the relative costs of implementing and using each ranking algorithm (component).
3) To determine specific file modifications necessary for conventional inverted file systems to implement particular ranking algorithms.

PROCEDURE

Briefly, the procedures followed in this study were:

1) To secure the use of a suitable data base. The specifics of the data base will be described later in this report.
2) To obtain the cooperation of a suitable user population. One hundred seventy three interest statements were acquired.
3) Review the term weighting schemes and select a representative group. Twenty one were finally selected.
4) Review the similarity measures both algebraically and by simulation to select a representative group. Twenty four were eventually chosen for inclusion in the study.
5) The statistical properties of the text were calculated to produce the weighting factors and similarity measures.
6) Characteristics of the data base were identified so that cost data could be acquired.
7) Programs were modified as necessary.
8) The data base was loaded and preliminary data was collected.
9) Intermediaries were trained to use the system. The intermediaries were kept blind of the system's ranking abilities.
10) Interest statements were obtained from users, and assigned to intermediaries.
11) The intermediaries translated the interest statements into Boolean queries. The interest statements were translated into the appropriate document representation.
12) Documents retrieved were merged and placed into a randomized order. This list was given to the user for relevance judgements.

13) Similarity values were computed between the query and the document. Documents were then rank ordered.
14) Ranking effectiveness scores were then calculated.
15) Cost data for the ranking algorithms were assembled.
16) Differences and patterns in the data were searched out.
17) Conclusions were drawn and appropriate post-hoc comparisons were performed.

REVIEW AND SELECTION OF TWs

This section reports on a review of the TWs found in IR literature and on the selection of a sample of TWs for inclusion in the experiment phase of the project. The experiment calls for the crossing of TW and SM in the environments of both DRs described above.

Since Luhn (1957) suggested that a term's frequency of occurrence in a document might be of value, in addition to its occurrence, about forty different TWs have been described in IR literature. The previously most comprehensive list of TWs was provided by Sager and Lockemann (1976).

Studies dealing with term weighting have had a variety of purposes, including recall or precision enhancement, selecting "good" index terms, term clustering, and ranking effectiveness. The TWs in these studies (see TW Bibliography) form the population from which the current work samples.

Certain types of TWs were excluded from consideration. These include manual weighting (e.g. Maron and Kuhns, 1960),

term classification schemes (e.g. Sparck Jones and Jackson, 1970), relevance weighting (e.g. Robertson, 1974), use of co-occurrence data (e.g. van Rijsbergen, 1977), and methods requiring complex estimation of distribution parameters (e.g. Harter, 1975). The first three types are not reasonably applicable to state-of-the-art automatic IR systems. The latter two have potential application, but are excluded from the present study because of their complexity and the effort required to implement and execute them on an operational system.

Even with the restrictions above, over thirty unique TWs were found. These measures differ on three basic dimensions:

1) The use of frequency information as opposed to binary (presence/absence) information about terms occurrences. TWs using frequency information are labelled "F" on Table 1 below.
2) Consideration of document length, (labelled "D").
3) Consideration of collection frequency, (labelled "C").

Table 1 contains a list of the major TWs considered. They vary as described above, as well as in the measures used to represent the component terms, operators connecting the terms, and scaling factors. The TWs in the literature have been based on theoretical grounds; the terms of the measure are related in order to represent theoretically defined relations. Yet the matrix of possible permutations of the terms is rather well filled in. Thus it seemed appropriate to define some new TWs to fill gaps

in the matrix. Also, some obvious simplifications are suggested.

Table 1 includes a reference for each TW, where appropriate, and the TW's form on the three dimensions. Other comments about each TW are reported, including reasons for the TW's inclusion (denoted by an *) or exclusion from the sample for experimentation. The sample was designed to allow generalizability to the population of TWs identified.

In addition to the specific results mentioned in Table 1, some general tendencies have been noted. Collection frequency has been successful, while the effects of within document frequency and document length have been ambiguous (Sparck Jones, 1973). Both TW specific and TW class differences were examined in the present study.

TERM WEIGHTINGS

TABLE 1

f_{in} = frequency of term i in document n.

t_n = number of types (unique terms) in document n.

d_i = number of postings of term i.

k_n = number of tokens (term occurrences) in document n. ($= \sum_{i=1}^{M} f_{in}$)

F_i = frequency of term i in data base. ($= \sum_{n=1}^{N} f_{in}$)

N = number of documents in data base.

M = number of terms in dictionary.

D = number of postings in data base. ($= \sum_{i=1}^{M} d_i$)

K = number of tokens in data base. ($= \sum_{n=1}^{N} \sum_{i=1}^{M} f_{in}$)

	FORMULA (W_{in} =)	REFERENCE	COMMENTS
* 1)	1	Sager (1976)	"Unweighted." Simplest and most commonly used method.
* 2)	$\frac{1}{t_n}$	Sager (1976)	D. Simplest consideration of document length.
3)	$\frac{1}{\log t_n}$	———	D. Obvious transformation to diminish effect of long documents.
4)	$10 - \text{intpt}(\frac{10\ t_n}{TOT\ MAX(t_n)})$	Sparck Jones & Bates (1977)	D. Integer formula. Where TOT(X)=next highest multiple of 10 above X.
5)	$2 - \frac{t_n}{MAX\ (t_n)}$	———	D. Non-integer version of #4.

TABLE 1

FORMULA (W_{in} =)	REFERENCE	COMMENTS
* 6) f_{in}	Sager (1976)	F. Simplest consideration of within document frequency
* 7) $\log f_{in}$	Sparck Jones (1973)	F. Diminishes effect of many occurrences in a document
* 8) $\frac{f_{in}}{k_n}$	Sager (1976)	FD. Simplest consideration of within document frequency and document length (using frequency information).
* 9) $\frac{f_{in}}{\log k_n}$	———	FD. Like #8 but diminishes impact of document length.
10) $f_{in}(10-\text{intpt}(\frac{10\ k_n}{TOT\ MAX(k_n)}))$	Sparck Jones & Bates (1977)	FD. Same as #4 but using frequency information
11) $f_{in}(2 - \frac{k_n}{MAX(k_n)})$	———	FD. Non-integer version of #10.
*12) $\frac{1}{d_i}$	Saeger (1976)	C. Simplest consideration of collection frequency.
13) $\frac{1}{\log d_i}$	———	C. Same as #12 but diminishes impact of high frequencies.
*14) $\log(\frac{N}{d_i})$	Saeger (1976) & Robertson (1974)	C. Based on the information content of a term about a document. Salton, Wong & Yang (1974) interpret this as $\log(\frac{N}{d_i}) + 1$.

TABLE 1

FORMULA (W_{in} =)	REFERENCE	COMMENTS
15) $\log \frac{N}{D}$	———	C. Percent of postings belonging to a term.
16) $G\log_2 N - G\log_2 d_i + 1$	Sparck Jones (1974)	C. Integer formula for #14. Where G(X)=M, where $2^{m-1} < X \leq 2^m$.
17) $\log \frac{K}{F_i}$	Sparck Jones & Bates (1976)	FC. #15 with frequency information.
* 18) $\frac{f_{in}}{F_i}$	Sparck Jones (1973)	FC. #12 with frequency information.
* 19) $f_{in} \frac{F_i}{d_i}$	Sager (1976)	FC. Available on IBM's STAIRS. Increases with higher ratio of occurrences per posting.
20) $\frac{f_{in}^2}{d_i}$	Sager (1976)	FC. Mixes levels, frequency information in numerator, binary in denominator.
21) $f_{in}^2 \cdot \frac{F_i}{d_i^2}$	Sager (1976)	FC. Like #19 but more sensitive to within-document frequency and also sensitive to high postings.

TABLE 1

FORMULA (W_{in} =)	REFERENCE	COMMENTS
* 22) $\frac{f_{in} d_i}{F_i - f_{in}}$	Sager (1976)	FC. Opposite effect of #19. Increases with fewer occurrences per posting.
23) $f_{in} \cdot \log \frac{N}{d_i}$	———	FC. #14 with frequency information, but mixed levels. Salton, Wong & Yang (1974) use $f_{in} \cdot \log(\frac{N}{d_i}) + 1$.
* 24) $f_{in} \cdot \log(\frac{K}{F_i})$	———	FC. #17 fully weighted, or less #23 without mixed levels.
* 25 $\frac{f_{in}}{\log F_i}$	———	FC. Like #18 but diminishes impact of collection frequency.
* 26) $\frac{1}{t_n d_i}$	Sparck Jones (1973)	DC. Naive combination. Simple consideration of document length and collection frequency.
* 27) $\frac{1}{\log(t_n d_i)}$	———	DC. Like #26 but diminishes effects of document length and collection frequency
28) $\frac{1}{t_n + d_i}$	———	DC. Not mentioned any where, but is simpler than #26 or #27.

TABLE 1

FORMULA (W_{in} =)	REFERENCE	COMMENTS
* 29) $\frac{1}{t_n} - \frac{d_i}{D}$	Sager (1976)	DC. Differences between term's role in document and in the collection.
30) $\frac{\frac{1}{t_n}}{\frac{d_i}{D}}$	Sager (1976)	DC. Reduces to $\frac{D}{t_n d_i}$ which is a linear transformation of #26.
* 31) $\frac{\frac{1}{t_n} - \frac{d_i}{D}}{\sqrt{\frac{d_i}{D}}}$	Sager (1976)	DC. Poisson Standard deviate (#37) converted for binary data.
* 32) $\frac{f_{in}}{k_n F_i}$	Sparck Jones	FDC. #26 with frequency information.
* 33) $\frac{f_{in}}{\log(k_n F_i)}$	———	FDC. #27 with frequency information.
34) $\frac{f_{in}}{k_n + F_i}$	———	FDC. #28 with frequency information
* 35) $\frac{f_{in}}{k_n} - \frac{F_i}{K}$	Edmundson & Wyllys (1961)	FDC. #29 with frequency information. Found effective in their study.

TABLE 1

FORMULA (W_{in} =)	REFERENCE	COMMENTS
36) $\dfrac{\frac{f_{in}}{k_n}}{\frac{F_i}{K}}$	Edmundson & Wyllys (1961)	FDC. #30 with frequency information. Found effective in their study. Is linear transformation of #32.
* 37) $\dfrac{\frac{f_{in}}{k_n} - \frac{F_i}{K}}{\sqrt{\frac{F_i}{K}}}$	Edmundson & Wyllys (1961)	FDC. Poisson Standard Deviate. Is #35 with difference standardized by an estimate of standard deviation.
38) $\dfrac{\frac{f_{in}}{k_n}}{\frac{f_{in}}{k_n} + \frac{F_i}{K}}$	Edmundson & Wyllys (1961)	FDC. Found ineffective in their study.
39) $\log\dfrac{\frac{f_{in}}{k_n}}{\frac{F_i}{K}}$	Edmundson & Wyllys (1961)	FDC. Found ineffective in their study.

REVIEW AND SELECTION OF SMs

SMs that potentially could be used to rank documents which have been retrieved by a Boolean query include any function which assigns a number to a pair of vectors based on their similarity. Selecting a representative sample of SMs presents more difficulties than does the selection of TWs. The main reason for this is that very few of the SMs advocated for use in IR have been created for the purpose of measuring the similarity of documents to queries, and very few actually have been used to rank order documents for output.

The SMs reviewed here are those that have been proposed or used for some IR activity, or are closely related to such measures in form or by reference. Many of the SMs come from the field of Numerical Taxonomy. Of these, some have been used for various purposes in IR, such as computing the similarities between terms or between documents for clustering.

Before selecting SMs for experimentation, it was useful to assemble a list of potential SMs. Of course, this list is not exhaustive, since the SMs come from such a diversity of areas. It was meant to be comprehensive in terms of those SMs mentioned in the context of IR, with certain restrictions. Certain types of SMs were excluded, namely those which are explicitly for measuring the similarity between groups (e.g. clusters of documents), measures requiring the changing of the nature of the attribute space (e.g. measures dependent on a factor analysis),

measures that place each item in a category rather than assign a score to each item, and iterative methods. This is not a major constraint since it leaves within our domain a rich population of SMs to which we can generalize.

A list of sixty-seven SMs is in Table 2. SMs marked by a + or an @ were used in the clustering analyses described below. SMs marked by an * were selected for the main experiment.

SIMILARITY MEASURES

TABLE 2

B = denotes binary measure
\+ = used in simulation 1
@ = used in simulation 2
* = used in main experiment
X'_i = 0 if $X_i > 0$, 1 otherwise
Y'_i = 0 if $Y_i > 0$, 1 otherwise

X_i = weight of term i in the document (X)
Y_i = weight of term i in the query (Y)
M = number of terms in dictionary
S_{xy} = denotes similarity measure

D_{xy} = denotes dissimilarity measures

	FORMULA	REFERENCE	COMMENTS
*@+ 1)	$S_{xy} = \frac{\Sigma X_i Y_i}{\sqrt{\Sigma X_i^2 \cdot \Sigma Y_i^2}}$	Torgerson (1958)	Cosine.
*@+ 2)	$S_{xy} = \Sigma X_i Y_i$	Overall and Klett (1972)	Vector or inner product
3)	$S_{xy} = \frac{1}{M} \Sigma X_i Y_i$	Overall and Klett (1972)	Mean Cross Product. Monotonic with #2.
*@+ 4)	$S_{xy} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma (X_i - \bar{X})^2 \cdot \Sigma (Y_i - \bar{Y})^2}}$	Sneath and Sokal (1973)	Pearson Product Moment Correlation.
B 5)	$S_{xy} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	Sneath and Sokal (1973)	Correlation for Binary Data. Equivalent to #4.

TABLE 2

FORMULA	REFERENCE	COMMENTS
*@+B 11) $S_{xy} = \frac{ad - bc}{M}$ $\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{M}\right)$	Jones & Curtis (1967) Maron & Kuhns (1960)	Maron and Kuhns'.
*@+B 12) $S_{xy} = \frac{ad - bc}{ad + bc}$ $\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i + \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}\right)$	Sneath and Sokal (1973)	Yule's. Numerator is determinant of 2 × 2 matrix.
B 13) $S_{xy} = \frac{ad - bc}{\sqrt{M(a+b)(a+c)}}$ $\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{\sqrt{M(\Sigma X_i Y_i + \Sigma X'_i Y_i)(\Sigma X_i Y_i + \Sigma X_i Y'_i)}}\right)$	Tague (1966)	Formula for converting binomial variable to standard normal form. Equivalent to #1.
+B 14) $S_{xy} = \frac{a+d-b-c}{M}$ $\left(= \frac{\Sigma X_i Y_i + \Sigma X'_i Y'_i - \Sigma X'_i Y_i - \Sigma X_i Y'_i}{M}\right)$	Sneath and Sokal (1973)	Hamann's. Found monotonic with #7 in binary simulation.

TABLE 2

FORMULA	REFERENCE	COMMENTS
+B 15) $S_{xy} = \frac{2a}{2a+b+c}$ $\left(= \frac{2\,\Sigma X_i Y_i}{\sqrt{\Sigma X_i^2} + \sqrt{\Sigma Y_i^2}} \right)$	Sneath and Sokal (1973)	Dice's SM
+B 16) $S_{xy} = \frac{a}{b+c}$ $\left(= \frac{\Sigma X_i Y_i}{\Sigma X_i' Y_i + \Sigma X_i Y_i'} \right)$	Lerman (1970)	Kulzynski's. Found monotonic with #15 in binary simulation.
+B 17) $S_{xy} = \frac{a}{a+2b+2c}$ $\left(= \frac{\Sigma X_i Y_i}{\Sigma X_i Y_i + 2\Sigma X_i' Y_i + 2\Sigma X_i Y_i'} \right)$	Lerman (1970)	Sokal and Sneath's. Found monotonic with #15 in binary simulation.
+B 18) $S_{xy} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$ $\left(= \frac{1}{2}\left(\frac{\Sigma X_i Y_i}{\Sigma X_i Y_i + \Sigma X_i' Y_i} + \frac{\Sigma X_i Y_i}{\Sigma X_i Y_i + \Sigma X_i Y_i'}\right) \right)$	Lerman (1970)	Kulzynski's. Arithmetic mean of shared percentage of X and Y.

TABLE 2

FORMULA	REFERENCE	COMMENTS
B 19) $S_{xy} = \frac{a}{\sqrt{(a+b)(a+c)}}$ $\left(= \frac{\Sigma X_i Y_i}{\sqrt{\Sigma X_i^2 \cdot \Sigma Y_i^2}} \right)$	Lerman (1970)	Geometric mean. Modified correlation coefficient. Equivalent to #1.
B 20) $D_{xy} = \frac{b + c}{2a + b + c}$	van Rijsbergen (1975)	Distance conversion of Dice's SM. (#15). Monotonic with #15.
B 21) $S_{xy} = \frac{a}{M}$ $\left(= \frac{\Sigma X_i Y_i}{M} \right)$	Lerman (1970)	Russell and Rao's. Equivalent to #2.
+B 22a) $S_{xy} = \frac{a}{a + c}$ $\left(= \frac{\Sigma X_i Y_i}{\Sigma X_i Y_i + \Sigma X_i Y'_i} \right)$	Jones and Curtis (1967)	Recall of Y for X. If $\lvert a+c \rvert < \lvert a+b \rvert$ then it is equivalent to #27. For binary data is monotonic with #1.

TABLE 2

FORMULA	REFERENCE	COMMENTS
+B 22b) $S_{xy} = \frac{a}{a+b}$ $\left(= \frac{\Sigma X_i Y_i}{\Sigma X_i Y_i + \Sigma X_i' Y_i} \right)$	Jones and Curtis (1967)	Recall of X for Y. If $\lvert a+c \rvert > \lvert a+b \rvert$ then it is equivalent to #27. For binary data is monotonic with #2.
B 23) $S_{xy} = (\frac{a}{a+b})^P (\frac{a}{a+c})^{1-P}$	Jones and Curtis (1967)	General form.
B 24) $S_{xy} = \frac{M a}{(a+b)(a+c)}$	Ball (1965)	Kochen and Wong. Equivalent to #1.
B 25) $S_{xy} = -\log \frac{M a}{(a+b)(a+c)}$	Ball (1965)	Abraham's. Equivalent to #1.
*@+B 26) $S_{xy} = \frac{M(a - \frac{1}{2})^2}{(a+b)(a+c)}$ $\left(= \frac{M(\Sigma X_i Y_i - \frac{1}{2})^2}{(\Sigma X_i Y_i + \Sigma X_i' Y_i)(\Sigma X_i Y_i + \Sigma X_i Y_i')} \right)$	Jones and Curtis (1967)	

TABLE 2

FORMULA	REFERENCE	COMMENTS
*@ 27) $S_{xy} = \frac{\Sigma_{min}(X_i, Y_i)}{min\ (\Sigma X_i, \Sigma Y_i)}$	Sager and Lockemann (1967)	Overlap. If $\|a+b\| > \|a+c\|$ then it is equivalent to #22a. If $\|a+b\| < \|a+c\|$ then it is equivalent to 22b.
28) $S_{xy} = \Sigma 1 - \frac{\|X_i - Y_i\|}{R_i}$	Sneath and Sokal (1973)	Gower, R_i = the maximum value of term i in any document.
*@+ 29) $S_{xy} = \frac{\Sigma(\frac{min\ (X_i, Y_i)}{max\ (X_i, Y_i)})}{N}$	Reitsma and Sagalyn (1968)	N = number of shared terms. For binary data is equivalent to #2.
*@+ 30) $S_{xy} = \frac{N(N-1)}{2}\ (\frac{\Sigma X_i Y_i}{M} - \frac{\Sigma X_i Y_i}{2M^2})$	Sager (1976)	Bennet and Spiegel. N = number of documents in collection.
* @+ 31) $S_{xy} = \frac{2M_{X.5(M)}^2 - \Sigma(X_i - Y_i)^2}{2M_{X.5(M)}^2 - \Sigma(X_i - Y_i)^2}$	Sneath and Sokal (1973)	Cattell's Pattern. Similarity. Found monotonic with #7 in binary simulation.

TABLE 2

FORMULA	REFERENCE	COMMENTS
*@+ 32) $S_{xy} = \frac{\Sigma(X_i + Y_i)\beta_i}{2N}$	Reitsma and Sagalyn (1968)	Average weight of shared terms. $\beta = 1$ if $X_i > 0$ and $Y_i > 0$, $= 0$ otherwise. N = number of shared terms. for binary vectors is equivalent to #2.
+B 33) $S_{xy} = \frac{\Sigma X_i Y_i}{\Sigma X_i + \Sigma Y_i - \Sigma X_i Y_i}$	Reitsma and Sagalyn (1968)	Parker-Rhodes Needham. Found monotonic with #15 for binary simulation. For binary data, is equivalent to #34.
B 34) $S_{xy} = \frac{a}{a + b + c}$	Sneath and Sokal (1973)	Jaccard's. Intersection divided by Union. For binary data is monotonic with #15.
B 35) $S_{xy} = \frac{a}{a + b + c}$	Tague (1966)	Doyle's. Equal to #34.
*@ 36) $S_{xy} = \frac{\Sigma X_i Y_i}{\Sigma X_i^2 + \Sigma Y_i^2 - \Sigma X_i Y_i}$	Sager and Lockemann (1976)	Tanimoto's. For binary data is equivalent to #33 and #34.

TABLE 2

	FORMULA	REFERENCE	COMMENTS
+ 37)	$S_{xy} = \frac{\sum^{N} X_i Y_i}{\sum^{N} X_i^2 + \sum^{N} Y_i^2 - \sum^{N} X_i Y_i}$	Reitsma and Sagalyn (1968)	Interpretation of #33 for weighted vectors, monotonic with #15.
+B 38)	$S_{xy} = \frac{ad - bc}{M \cdot \max(a+b,\ a+c)}$	Kuhns (1965)	Rectangular distance above independence. Found monotonic with #1 in binary simulation.
	$\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{M \cdot \max(\Sigma X_i Y_i + \Sigma X'_i Y_i,\ \Sigma X_i Y_i + \Sigma X_i Y'_i)} \right)$		
+B 39)	$S_{xy} = \frac{2(ad - bc)}{M^2}$	Kuhns (1965)	Separation above independence. Monotonic with #11.
	$\left(= \frac{2(\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i)}{M^2} \right)$		
*@+B 40)	$S_{xy} = \frac{2(ad - bc)}{M(2a + b + c)}$	Kuhns (1965)	Coefficient of the Artihmetic Mean
	$\left(= \frac{2(\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i)}{M(2\Sigma X_i Y_i + \Sigma X'_i Y_i + \Sigma X_i Y'_i)} \right)$		

TABLE 2

FORMULA	REFERENCE	COMMENTS
B 41) $S_{xy} = \frac{2(ad - bc)}{M\sqrt{(a+b)(a+c)}}$	Kuhns (1965)	Angle between vectors above independence. For binary data is monotonic with #1.
$\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{M\sqrt{(\Sigma X_i Y_i + \Sigma X'_i Y_i)(\Sigma X_i Y_i + \Sigma X_i Y'_i)}} \right)$		
+B 42) $S_{xy} = \frac{ad - bc}{M \cdot \min(a+b - \frac{(a+b)^2}{M}, a+c - \frac{(a+c)^2}{M})}$	Kuhns (1965)	Probability difference 2 above independence. Found monotonic with #11 in binary simulation.
$\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{M \cdot \min(\Sigma X_i Y_i + \Sigma X'_i Y_i - \frac{(\Sigma X_i Y_i + \Sigma X'_i Y_i)^2}{M}, \Sigma X_i Y_i + \Sigma X_i Y'_i - \frac{(\Sigma X_i Y_i + \Sigma X_i Y'_i)^2}{M})} \right)$		
+B 43) $S_{xy} = \frac{ad - bc}{M \cdot \max(a+b - \frac{(a+b)^2}{M}, a+c - \frac{(a+c)^2}{M})}$	Kuhns (1965)	Probability difference 1 above dependence. Found monotonic with #7 in binary simulation.
$\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{M \cdot \max(\Sigma X_i Y_i + \Sigma X'_i Y_i - \frac{(\Sigma X_i Y_i + \Sigma X'_i Y_i)^2}{M}, \Sigma X_i Y_i + \Sigma X_i Y'_i - \frac{(\Sigma X_i Y_i + \Sigma X_i Y'_i)^2}{M})} \right)$		

TABLE 2

FORMULA	REFERENCE	COMMENTS
*@+B 44) $S_{xy} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$	Kuhns (1965)	Yule's coefficient of colligation. Found very similar to #12 in both simulations.
$\left(= \frac{\sqrt{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i} - \sqrt{\Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}}{\sqrt{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i} + \sqrt{\Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}} \right)$		
+B 45) $S_{xy} = \frac{ad - bc}{M \cdot \min(a + b,\ a + c)}$	Kuhns (1965)	Conditional probability above independence. Found monotonic with #11 in binary simulation.
$\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{M \cdot \min(\Sigma X_i Y_i + \Sigma X'_i Y_i,\ \Sigma X_i Y_i + \Sigma X_i Y'_i)} \right)$		
*@B 46) $S_{xy} = \frac{ad - bc}{M(1 - \frac{a}{(a+b)(a+c)}(2a+b+c - \frac{(a+b)(a+c)}{M})}$	Kuhns (1965)	Proportion of overlap above independence.
$\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{M(1 - \frac{\Sigma X_i Y_i}{(\Sigma X_i Y_i + \Sigma X'_i Y_i)(\Sigma X_i Y_i + \Sigma X_i Y'_i)}(2\Sigma X_i Y_i + \Sigma X'_i Y_i + \Sigma X_i Y'_i - \frac{(\Sigma X_i Y_i + \Sigma X'_i Y_i)(\Sigma X_i Y_i + \Sigma X_i Y'_i)}{M})} \right)$		

TABLE 2

	FORMULA	REFERENCE	COMMENTS
B 47)	$S_{xy} = \frac{ad - bc}{(a+b)(a+c)}$ $\left(= \frac{\Sigma X_i Y_i \cdot \Sigma X'_i Y'_i - \Sigma X'_i Y_i \cdot \Sigma X_i Y'_i}{(\Sigma X_i Y_i + \Sigma X'_i Y_i)(\Sigma X_i Y_i + \Sigma X_i Y'_i)} \right)$	Kuhns (1965)	Index of Independence. For binary data is monotonic with #1.
+B 48)	$S_{xy} = M\ln M + a\ln a + b\ln b + c\ln c + d\ln d - (a+b)\ln(a+b) - (a+c)\ln(a+c) - (b+d)\ln(b+d) - (c+d)\ln(c+d)$	Sneath and Sokal (1973)	Mutual Information of X and Y. I (X;Y) equals I (X) + I (Y) - I (X,Y) where I (X) = information on X, I (X,Y) = joint information on X and Y. Not readily applicable to weighted vectors.
+B 49)	$S_{xy} = [M\ln M + a\ln a + b\ln b + c\ln c + d\ln d - (a+b)\ln(a+b) - (a+c)\ln(a+c) - (b+d)\ln(b+d) - (c+d)\ln(c+d)] / [M\ln M - a\ln a - b\ln b - c\ln c - d\ln d]$	Orlocci (1969)	Ratio of mutual to joint information. Equals I(X;Y)/I(X,Y). Not readily applicable to weighted vectors.
+B 50)	$S_{xy} = \sqrt{1 - (1 - \frac{I(X;Y)}{I(X,Y)})^2}$	Orlocci (1967)	Rajski's Coherence Coefficient. Not readily applicable to weighted vectors.

TABLE 2

FORMULA	REFERENCE	COMMENTS
+ 51) $S_{xy} = \dfrac{\sum_{k=1}^{N} \left(\sum_{i=1}^{n} X_{ik} \cdot \sum_{j=1}^{m} Y_{jk} \right)}{\sum_{k=1}^{N} \left(\sum_{i=1}^{n} (X_{ik})^2 \cdot \sum_{k=1}^{M} (Y_{jk})^2 \right)}$	_____	Where: N = number of documents in collection. n = number of terms in document (X). m = number of terms in Query (Y). X_{ik} = frequency of document (X)'s ith term in document K. Angle between average term of document and average term of query over space defined by documents. Requires excessive computation.
+~~0~~+ 52) $S_{xy} = \dfrac{100\ \Sigma B(X_i,Y_i)}{\log P(i)} + \dfrac{10\ \Sigma X_i Y_i}{\log P(i)} + \dfrac{\Sigma X_i Y_i}{\log P(i) \cdot \sqrt{\Sigma X_i^2}}$	_____	Where $\beta = 1$ if $X_i > 0 < Y_i$ = 0 otherwise. P(i) = number of postings of term i. For binary data is equivalent to #18.

TABLE 2

	FORMULA	REFERENCE	COMMENTS
+B 53)	$S_{xy} = \frac{M(\lvert Ma - (a+b)(a+c)\rvert - \frac{M}{2})^2}{(a+b)(a+c)(b+d)(c+d)}$	Jones and Curtis (1967) Reitsma and Sagalyn (1968)	Stile's. For binary data is equivalent to #5 and #4.
	$\left(= \frac{M(\lvert M\Sigma X_i Y_i - (\Sigma X_i Y_i + \Sigma X_i' Y_i)(\Sigma X_i Y_i + \Sigma X_i Y_i')\rvert - \frac{M}{2})^2}{(\Sigma X_i Y_i + \Sigma X_i' Y_i)(\Sigma X_i Y_i + \Sigma X_i Y_i')(\Sigma X_i' Y_i + \Sigma X_i' Y_i')(\Sigma X_i Y_i' + \Sigma X_i' Y_i')} \right)$		
@ 54)	$S_{xy} = \frac{4M(\lvert \frac{N\Sigma X_i Y_i}{144} - \frac{\Sigma X_i^2 \cdot \Sigma Y_i^2}{144}\rvert - \frac{M}{2})^2}{(\frac{\Sigma X_i^2}{144})(\frac{\Sigma Y_i^2}{144})(4M - \frac{\Sigma X_i^2}{144})(4M - \frac{\Sigma Y_i^2}{144})}$	Reitsma and Sagalyn (1968)	Their interpretation of #53 for weighted vectors.
@ 55)	$D_{xy} = \sqrt{\Sigma(X_i - Y_i)^2}$	Sneath and Sokal (1973)	Euclidean distance. Minkowski 2. Equivalent to $\Sigma X_i^2 + \Sigma Y_i^2 - 2\sqrt{\Sigma X_i^2 \cdot \Sigma X_i^2} \cdot \cos_{xy}$ and $\Sigma X_i^2 + \Sigma Y_i^2 - 2\Sigma X_i Y_i$. Found monotonic with #31 in both simulations.

TABLE 2

	FORMULA	REFERENCE	COMMENTS
*@ 56)	$D_{xy} = \Sigma\lvert X_i - Y_i\rvert$	Sneath and Sokal (1973)	City Block distance. Minkowski 1. For binary data is equivalent to #55.
57)	$D_{xy} = \frac{1}{M}\Sigma\lvert X_i - Y_i\rvert$	Sneath and Sokal (1973)	Mean Character Difference. Equivalent to #56.
*@+ 58)	$D_{xy} = \frac{1}{M}\Sigma(X_i - Y_i)$	Sneath and Sokal (1973)	Average Distance.
59)	$D_{xy} = \sqrt{\frac{\Sigma(X_i - Y_i)^2}{M}}$	Sneath and Sokal (1973)	Euclidean Distance Average. Equivalent to #55.
60)	$D_{xy} = \Sigma W_i(X_i - Y_i)^2$	Cormack (1971)	General Euclidean Distance Form. W can equal 1, or $\frac{1}{\sigma_i^2}$, or $\frac{1}{\text{Max}(X_{ij} - Y_{ij})}$ for all j.
61)	$D_{xy} = [\Sigma(X_i - Y_i)^P]^{1/P}$	Sneath and Sokal (1973)	General Minkowski Form.

TABLE 2

	FORMULA	REFERENCE	COMMENTS
+ 62)	$S_{xy} = \frac{\sqrt{2M} + \sqrt{\Sigma(X_i - Y_i)^2}}{\sqrt{2M} + \sqrt{\Sigma(X_i - Y_i)^2}}$	Cormack (1971)	Coefficient of Nearness. Equivalent to #55. Found monotonic with #7 in binary simulation.
*@+ 63)	$D_{xy} = \Sigma\left(\frac{\lvert X_i - Y_i \rvert}{(X_i + Y_i)}\right)$	Sneath and Sokal (1973)	Canberra Distance For binary is equivalent to #7.
*@+ 64)	$D_{xy} = \frac{1}{M}\Sigma\left(\frac{X_i - Y_i}{X_i + Y_i}\right)^2$	Sneath and Sokal (1973)	Coefficient of Divergence. For binary data is monotonic with #7 and #63.
+B 65	$D_{xy} = \frac{b + c}{2a + b + c}$	Sneath and Sokal (1973)	Nonmetric Coefficient. Equal to #20. For binary data is equivalent to #20, #15.
+B 66)	$S_{xy} = \frac{a + d - b - c}{a + d + b + c}$	Lerman (1970)	Harman's. Equal to #14.
	$\left(= \frac{\Sigma X_i Y_i + \Sigma X'_i Y'_i - \Sigma X'_i Y_i - \Sigma X_i Y'_i}{\Sigma X_i Y_i + \Sigma X'_i Y'_i + \Sigma X'_i Y_i + \Sigma X_i Y'_i}\right)$		
+B 67)	$S_{xy} = \frac{a + d}{a + 2b + 2c + d}$	Sneath and Sokal (1973)	Rogers and Tanimoto's. Monotonic with #9 and #7.
	$\left(= \frac{\Sigma X_i Y_i + \Sigma X'_i Y'_i}{\Sigma X_i Y_i + 2\Sigma X'_i Y_i + 2\Sigma X_i Y'_i + \Sigma X'_i Y'_i}\right)$		

SMs marked by a B are designed for use on binary vectors. Binary measures are described in terms of the two-by-two table shown below. (Table 3).

		X = 1	X = 0	
Y	1	a	b	ΣY
	0	c	d	ΣY'
		ΣX	ΣX'	

where,

$$a = \sum_{i=1}^{M} X_i Y_i$$

$$b = \sum_{i=1}^{M} X_i' Y_i$$

$$c = \sum_{i=1}^{M} X_i Y_i'$$

$$d = \sum_{i=1}^{M} X_i' Y_i'$$

X_i = 1 if Doc x contains term i
0 otherwise

X_i' = 1 if X_i = 0
0 if X_i = 1

Y_i = 1 if Query Y contains term i
0 otherwise

Y_i' = 1 if Y_i = 0
0 if Y_i = 1

M = number of terms in index

Below many of the binary measures appears a generalized version intended for use on weighted vectors. The translation is not always obvious (cf. Reitsma and Sagalyn, 1968). In each case the weighted version was constructed so that the binary measure is a special case of the weighted version and so that the characteristics of similarity being measured are preserved as much as possible. The first point means that when applied to binary vectors, the binary SM and its generalized counterpart produce equivalent values. The second point refers to the intended behavior of the SM. For example, a + b + c + d could be interpreted as a constant, M, or as an additive function of the vector lengths, $|X|$ and $|Y|$. In such cases, an attempt was made to preserve the intention of the binary SM.

A reference is also given for each SM. This reference is either to the measure's introduction or to a discussion or analysis of the measure. Preference was given to more accessible sources.

The SMs listed in Table 2 may be divided into four types following the typology of Sneath and Sokal (1973). "Association coefficients" work with qualitative data and measure some variant of the amount of agreement between the two items. The binary SMs discussed above fall under this heading.

"Correlation coefficients" covers such measures as the Pearson Product Moment Correlation and the cosine correlation. Such SMs measure proportionality and departure from independence.

"Distance measures" obviously measure dissimilarity, smaller values indicating greater similarity. The notion of distance implies a space in which distance is measured. The distance measures described here do not necessarily obey the rules of Euclidean spaces. Distance SMs are denoted by a "Dxy" instead of an "Sxy" on Table 2.

"Probability coefficients" treat the likelihood of agreements as well as the presence or amount of agreement as considered by the other measures. Such SMs include information theoretic values.

After compiling this list, it was desirable to narrow it down to a group of approximately twenty for the main experiment.

Comparison by inspection and simple algebraic manipulation enabled some reduction of the list. Some formulas were found to be identical, having been described in the literature by different names or different terminology (see example No. 34 and No. 35). Other pairs were found to have joint monotonicity, (i.e. the rankings they produce are identical). SMs No. 3 and No. 2 are examples of this. Many relationships among SMs were more subtle. Two simulation studies were run to identify, a) pairs of SMs which are monotonic even though this might not be obvious through algebraic comparison, and b) clusters of SMs which tend to produce similar rank orderings of documents.

The first simulation used binary information about the presence of terms in documents (i.e. TW No. 1). This limited the conclusions that could be drawn from this study, but simplified it considerably. An aim of this simulation was to decrease the number of SMs considered to be unique. Specifically, if a measure which was intended for use on binary data was found to produce identical or similar rank orderings to another binary SM or to an SM intended for weighted vectors, then a binary SM could be dropped from further analysis. In the first case, there is not sufficient reason for translating an SM beyond its intended application if it is not even unique within its own sphere. In the second case, the binary SM is found to be a special case of a more general SM.

The simulated products of these two simulation studies were the document orderings that would be produced by the various

SMs. For the first simulation, an artificial binary term-by-document matrix was constructed. This matrix consisted of a simulated sample of fifty "documents" and fifty "terms". The presence/absence data was created using random numbers. The numbers were drawn from distributions approximately the same as those observed in the free text DR of the CIJE data base. Thus, the parameters describing the indexing breadth and depth of the simulated sample approximate those of the CIJE data base.

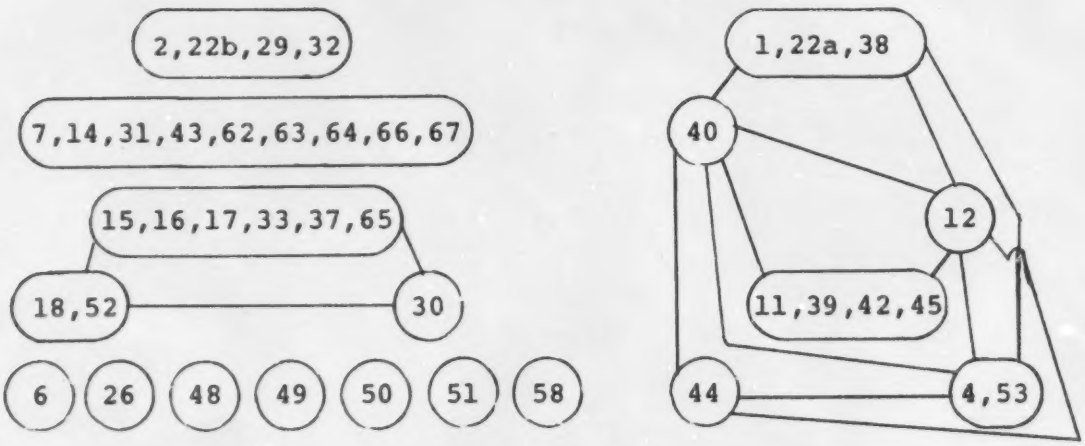
The first simulation consisted of submitting "queries" (artificially constructed to have the same distribution of words as would be found normally), and determining the rank orderings of the fifty documents that would be produced by the various SMs. These orderings were compared by computing rank order correlations between the ordered lists produced by each pair of SMs. These correlations were averaged over fifteen simulated queries.

The pattern of correlations can be seen in the graph below, (Figure 4).

CORRELATIONS BETWEEN SMs

BASED ON BINARY SIMULATION OF 15 QUERIES

FIGURE 4



* Each set of circled number(s) represents one unique SM. Edges represent correlations > .7.

These correlations depart slightly from the central aim of the study. They are based on a ranking of all fifty (simulated) sample documents, whereas the main study is only concerned with ranking documents which fulfill the logic of a query. In practice this means ranking documents that at least have one term in common with the query. Some SMs differed in this simulation only on documents which would not normally be retrieved. In these cases, the SMs were grouped as equivalent. The effect of this difference is that the correlations are generally lower than they would be under retrieved set ranking.

Some SMs (Nos. 6, 26, 48, 49, 50, 51, 58) had conspicuously low (near zero) correlations with all other SMs. As a result of the analyses to this point, twenty-nine unique SM types may be described. These are listed in Table 3. Some SMs listed as unique were found to have very high or perfect correlations with other SMs in the binary simulation. SMs designed for weighted vectors were retained because they might differ more when weights are used. Note that SMs such as Nos. 40, 44, 12 and 4 (5, 53) are considered unique at this point, despite high correlations, because of potential differences on weighted vectors.

UNIQUE SMs AFTER BINARY SIMULATION

TABLE 3

SM No.	(Others Monotonic With It)
1	(13, 19, 22, 24, 25, 38, 41, 47)
2	(3, 21, 22b)
4	(5, 53)
6	
7	(8, 9, 10, 14, 43, 66, 67)
11	(39, 42, 45)
12	
26	
27	
28	
29	
30	
* 31	
32	(15, 16, 17, 20, 33, 34, 35, 37, 65)
40	
44	
46	
48	
49	
50	
51	
52	(18)
54	
55	(59, 62)
56	(57)
58	
63	
64	

* Found monotonic with No. 55 in weighted simulation.

In Table 3, the SMs presented outside the parentheses are either unique, or represent several SMs which are equivalent, but, as a group are unique from other SMS. The SMs representing groups were selected on the basis of generality (general over Boolean) and computational simplicity.

A second simulation study was performed using frequency information (TW No. 6) instead of binary information (TW No. 1). The frequency information was added to the fifty-by-fifty matrix used previously in such a way so as to model the term frequency distributions of the CIJE data base. Again, the correlation between the document orderings created by the SMs were averaged over fifteen simulated queries.

This simulation study looked for relationships among the heretofore unique SMs. The correlations obtained in this study were much lower; very few were as great as 0.4. Based on these low correlations, we may conclude that the document orderings produced by these SMs were different from each other.

Two relationships were observed in the second simulation which helped select a sample for later experimentation. SM No. 5, which was not included in the first simulation, was found equivalent to No. 31. Also, SMs No. 12 and No. 44 again displayed a very strong correlation ($> .9$).

This left twenty eight SMs in the sample. For reasons such as high correlation or computational complexity, some of these SMs were excluded from the sample used for further experimentation, bringing the sample size to twenty-four. Inclusion and reasons for exclusion are noted in Table 2.

DESCRIPTION AND LOADING OF THE DATA BASE

INTRODUCTION

This section describes the data base and each of the document representations used for this study. The data base is a subset of the ERIC CIJE (Current Index to Journals in Education). The document representations chosen were (1) terms from the title and annotation, and (2) the ERIC descriptors for each document. These representations were selected as representatives of those available in commercial data bases.

Additionally, this report presents a comparison of the CPU use and storage requirements for the loading of the data base. This is the computer costs for construction of the dictionary to make the system available online. None of the computer or labor costs involved in the construction of the data base are included.

Description of Data Base

The data base for this project consisted of 10885 records from CIJE. The selected records are from four clearinghouses: Tests, Measurement and Evaluation (TM), Information Resources (IR), Educational Management (EA), and Teacher Education (SP). At the time of acquisition (August

1978) these were the most recent records availabe from ERIC in each of the clearinghouses.

The distribution of the records among the clearinghouses is:

	% of Total	Records
EA	31.8	3461
IR	26.9	2928
SP	28.8	3135
TM	12.5	1361

TABLE 4

RECORDS IN DATA BASE BY CLEARINGHOUSE

These reflect the proportion of the ERIC CIJE data base developed by each of the clearinghouses. Records were selected by identifying those developed by the four clearinghouses over the period of the previous 24 months. No other selection criteria were used.

Each of the document representations, controlled descriptors and free text, was used to create a separate inverted file. The free terms, from title and abstracts, were compared to a stop list containing about 150 common terms and then the remaining terms were stemmed (TARS, 1976). The stemming algorithm reduces the number of free terms. This in turn reduces the need to identify all work variations for retrieval. Controlled descriptors were used as developed by

the ERIC professionals. For purposes of system compatability imbedded blanks were removed. Each controlled descriptor was truncated at 24 characters. This insured the uniqueness of each descriptor. This same process was conducted at the time of the search and was transparent to the intermediary.

Statistics describing the characteristics of the data base for the controlled and free terms are presented in Table 5.

	Controlled	Free
Average number of terms in a record	6.45	20.39
Average number of unique terms in a record	6.45	16.77
Average number of postings in a term	18.21	17.62
Number of unique terms	3855	10361

TABLE 5

DESCRIPTION OF DATA BASE BY REPRESENTATION

Construction of the Inverted Files

As noted earlier, an inverted file was constructed for both the controlled descriptors and free text document representations. These files were constructed using SIRE (Syracuse Information Retrieval System). SIRE (McGill et al, 1976) was developed at the School of Information Studies for experimental use. This section will explain the process used in constructing the inverted files.

SIRE uses a three-step process in building the inverted file.

1. Each record is processed sequentially, producing a dictionary of the terms in that record. The output from this step is a file in which each record consists of a term, frequency of the term in document, and document number. Other small files are produced at this point which contain pointers and document length. The SIRE program which accomplishes this is READIN (See Figure 5).

2. The next process is to sort the file produced by READIN into alphabetical order on the term field. This step is accomplished by two programs, SORT and MERGE. (See Figure 6).

3. The final step is to use the sorted file to produce an inverted file. The program to accomplish this is MAKDIC. (See Figure 7).

These processes were used to construct the inverted file for the free terms. Modification of the process was required for the controlled terms since SIRE was designed to handle individual terms with a length of up to 12 characters. Since controlled terms were often phrases or combinations of words, more than 12 characters were needed to assure that each controlled term had a unique representation. This required inserting a new step between READIN and MERGE/SORT which converted the controlled terms into codes and constructed a

conversion table. This conversion program is to be called CONNUM.

Comparison of Construction Times

The CPU usage and total cost for each of the programs used in the construction of the inverted file are provided in Table 6. These figures are most useful for comparison of relative costs of the two document representations. Actual costs are dependent on the particular configuration and characteristics of the computer installation. In particular, the cost of this SORT procedure is inflated since the SORT and MERGE were written locally and are less efficient than commercially available packages. All programs are written in SAIL (Stanford Artificial Intelligence Language), an ALGOL-60 variant, on a DEC-10.

The comparison of the controlled and free text shows that the controlled is less expensive to build and store. Specifically, only 60% of the CPU time and 90% of the space to store the dictionary were used for the controlled descriptors as compared to the free text.

The controlled vocabulary is a less expensive representation, in terms of computer usage, than is the free text. Computer costs for building and maintaining a document representation based on free text would be greater than those for a document representation based on controlled terms.

	Controlled	Free	
READIN	210	460.33	
CONNUM	267	- -	CPU time in seconds.
SORT	264.58	667.16	
MERGE	44.69	169.87	
MAKDIC	7.66	16.97	
Total	793.93	1314.33	

TABLE 6

PROCESSING TIMES FOR FILE CONSTRUCTION

	File Sizes (36 Bit Words)		
	Variable	Fixed	Total
Controlled	85,632	1,105,430	1,101,062
Free	213,632	1,105,430	1,319,062

TABLE 7

COMPARISON OF FILE SIZES

The fixed file size includes pointer files and the CIJE data base source. The variable file size is the inverted file, which will be different for the different representations.

FIGURE 5. OUTPUT FROM READIN

Term	Freq	Document 1
.	.	.
.	.	.
.	.	.
Term	Freq	Document 2
.	.	.
.	.	.
.	.	.
Term	Freq	Document 3
.	.	.
.	.	.
.	.	.
Term	Freq	Document N

File is in order of increasing document number. Within a document each term occurs only once and records are in alphabetic order on terms.

FIGURE 6. OUTPUT FROM MERGE/SORT

Term	Freq	Document No.
.	.	.
.	.	.
.	.	.
.	.	.
Term	Freq	Document No.

File from Figure 1 has been sorted into increasing order on term.

FIGURE 7. OUTPUT FROM MAKDIC

Index

Term	No.of Terms	Pointer
.	.	.
.	.	.
.	.	.

Inverted File

Term	Posting	Pointer
.	.	.
.	.	.
Term	Posting	Pointer
Doc. No.		Freq.
.		.
.		.
.		.

.	.
.	.
.	.

File is produced from file in Figure 6 and is the inverted file which SIRE used for retrieval.

COLLECTING INTEREST STATEMENTS

The Computer Index to Journals in Education is a data base with a broad group of potential users. By selecting the clearinghouses on Tests, Measurements, and Evaluation, Educational Management, Information Resources, and Teacher Education, the group of potential users was narrowed considerably. The final users were from Syracuse University, Cornell University, and the local Syracuse geographic area. They included students, faculty, and local professionals. In order to assure the users of complete anonymity, no specific demographic data were collected.

Users were individuals with actual information requirements. A psuedo-service was established and appropriate announcements were made of its availability in classrooms, through mailings, and by word of mouth. A copy of the announcement flier is included in Appendix B. Information request statements were collected on the request forms included in Appendix B. The forms were acquired from 25 October, 1978 through 15 February, 1979. A total of one hundred seventy-three information request statements were received, searched, and sent back to the user for relevance judgments. One hundred forty were returned with completed relevance judgments for a response rate of 80.9%.

The study required a comparison of representations and a measure of the system's ability to rank relevant documents

within a retrieved set. If a specific retrieval set contains only relevant documents, then one is unable to measure an algorithm's ability to place relevant documents before non-relevant documents. The same is true for a search which retrieves no relevant documents. Thus for the purpose of this study, in order for a query to be useable, one relevant and one non-relevant document had to be retrieved from each representation. Of the 140 searches which were returned with relevance judgments, 68 had at least one relevant and one non-relevant document in each representation. Thus, 48.5% of the completed searches with relevance judgments were useable for the study of ranking algorithms.

A sample of the returned output along with the instructions to users for relevance judgments is included in Appendix B.

INTERMEDIARIES AND SEARCHING

The three intermediaries were selected because they are professional searchers. At the beginning of the study, each intermediary had been searching the entire ERIC data base for at least one year. The intermediaries were given a brief training session on the use of the SIRE system and a one page description of the appropriate commands for this study (Appendix A). These instructions did not include any description of SIRE's ranking capability nor any of the natural language features for searching. Thus the searchers were kept

unaware of the goal of the study and were unable to use techniques which might have contaminated the study. Searchers were instructed to perform high recall searches.

Each information request form was duplicated so that it could be sent to one intermediary for searching in the controlled vocabulary and to one for searching in the free term vocabulary. The information requests were assigned randomly to the intermediaries. Each intermediary was instructed to conduct a search in the appropriate vocabulary, controlled or free. By random selection, sixty eight information requests were searched by the same intermediary in both the free and the controlled vocabulary. Each of the remaining one hundred five was searched by different intermediaries in each of the vocabularies. Twenty three queries in the free representation and twenty seven queries in the controlled representation were randomly selected for reliability checks. Each of these information request statements was submitted to all of the intermediaries for a consistency check on performance. The documents retrieved by the originally designated intermediary were returned to the user for relevance judgments. Documents retrieved by the remaining two intermediaries were used for an examination of the overlap of the document sets.

The output forms returned to users for relevance judgments were limited to fifty documents. The thirty three queries that retrieved more than fifty documents were reduced to fifty by randomly selecting from those documents in the

full retrieved set. The retrieved output was placed in a random order prior to return to the user to control for any order effect. A statistical test for order effect was conducted. For all practical purposes no correlation was found. The correlation between position on the list and a positive relevance judgment was .01.

In most cases retrieved output was hand delivered to the user. Some output was mailed and in some instances, it was picked up directly from the office. A form letter was developed requesting the return of the evaluated output after a reasonable period of time had passed and the relevance judgment had not been returned. This is included in Appendix B. This proved to be an effective means of increasing the return. Sixty two reminder letters were sent out, resulting in the return of thirty nine evaluated output forms or 62.9% of the reminders resulted in returned forms.

Users were asked to evaluate each retrieved reference on a scale of 1 to 4, where 1 indicated direct relevance to the information request, and 4 indicated no relevance to the information request.

The data required for this study were the output documents and the associated relevance judgments. Each information request was kept as an independent unit. For each request, data was captured indicating the documents retrieved, the relevant documents, the non-relevant documents, and any documents which

were not returned to the users. Relevance information was stored as the original 1 to 4 values assigned by the user. However, for the study this information was dichotomized to indicate the relevance and non-relevance of a document to a particular request.

QUERY PROCESSING

Query processing included the acquisition of interest statements, clerical procedures prior to sending the information request to intermediaries, the actual processing (searching) by the intermediaries, computer programs to collect and rearrange the references, preparation of the output and judgment information, delivery of the references to the users for evaluation, and return to the project office for input. The results of this entire process were the data required for an analysis of the ranking algorithms. Thus, the key factors are the requests and the characteristics of the output developed from the requests.

To characterize the requests and the output, one begins with the form of the requests. Each user was requested to submit a two or three sentence statement in plain English describing their information need. In fact, most requests were two or three sentences, with the outliers ranging from a few descriptive words (as if selected from a controlled vocabulary) to as many as ten sentences. Each information

request form was delivered to the appropriate randomly selected intermediary. There was no direct contact between the intermediary and the user. Each request was processed according to the previous instructions and when the intermediary was satisfied with the output, the search was terminated. No controls were put on the length of the search, either in time or number of commands.

The data in Table 8 shows the operational characteristics of the search process. The average number of references retrieved in response to an information request was 18.7. This ranged from searches which retrieved 0 items to those which retrieved 170 items.

In the free representation, 17 of the 55 useable queries retrieved more than 50 documents or 63.6% of the queries retrieved 50 or less documents. 80% retrieved 55 or fewer documents and 89.1% retrieved 72 or fewer documents. The controlled representation shows there were 68 useable queries with 35 or 51.5% retrieving 50 or fewer documents. 80.9% retrieved 72 or fewer documents and 89.7% retrieved 103 or less documents. There is a major difference between one searcher and the other two searchers in terms of the number of items retrieved. Searchers A and B retrieved an average of 24.3 documents per query while Searcher C retrieved an average of 7.3. Thus Searcher C retrieved 69.9% fewer documents on the average than either Searcher A or B.

		Number of Searches	Average Retrieved	Difference Between Controlled and Free	Average Relevant	Difference in Relevant Between Controlled and Free	Precision	Difference in Precision Between Controlled and Free
A	Free	27	25.1	+5.3	7.6	+0.6	.37	-.04
	Controlled	25	19.8		7.0		.41	
	Total	52	22.6		7.3		.39	
B	Free	28	29	+6.3	8.4	-0.7	.34	-.21
	Controlled	26	22.7		9.1		.55	
	Total	54	26		8.7		.42	
C	Free	25	7.3	0.0	2.7	-1.0	.40	-.24
	Controlled	27	7.3		3.7		.54	
	Total	52	7.3		3.2		.47	

TABLE 8

SUMMARY OF SEARCH CHARACTERISTICS

On the other hand, the total precision performance figures vary by only 88%. The conclusion is that there is a significant difference in the documents retrieved by the intermediaries, but the difference in performance measures is slight.

The searchers differed significantly in their performance when examined by representation. Searchers A and B clearly retrieved more documents in the free representation than in the controlled. Searcher C performed identically in both representations. However, the number of relevant retrieved documents shows that searchers B and C were able to use the controlled representation more effectively.

The lack of agreement among documents retrieved across across representations and searchers clearly does not affect the precision achieved by the intermediaries. Precision ranged from .34 to .54. Within the free representation, the observed precision ranged from .34 to .40. Within the controlled representation precision ranged from .41 to 54. The controll representation provided consistently better precision in this study with a searcher difference between representations ranging from .04 to .21. One a priori factor constant among the intermediaries is their previous experience with the ERIC controlled vocabulary. This may influence the direction and/or the magnitude of the observed data.

To examine differences in the documents retrieved by searchers, an overlap study was conducted using the 33 queries identified for the reliability data. The results of this study are shown in Table 9.

		SEARCHER	
		Same	Different
Representation	Same	--	9%
Free and Controlled	Different	14%	5%

TABLE 9

OVERLAP PERCENTAGES

The observed overlap is very small. In fact, these figures could be explained by chance. By chance is meant that the representations are independent with respect to their descriptions of relevant documents and their descriptions of non-relevant documents. That is, both representations may perform similarly in discriminating the relevant from the non-relevant documents. However, within relevant or non-relevant subsets there is no relationship between the way the documents are described. It is as if retrieval using either representation was done by randomly sampling from the same sets of relevant and non-relevant documents. Thus different sets of documents (but the same relevant/non-relevant percentages) are retrieved by each. Another explanation may be that the representations are systematically different. That is, a searcher knowing that one or the other representation is being used, will systematically retrieve different documents. The data from this study does not allow for an in-depth examination of these findings. Further data focused on this topic are required for a complete understanding of these preliminary findings.

The complete records of searches were retained, thus allowing exploratory analyses to be perused. Two of the more interesting features were discovered by looking for factors which correlated with precision. Neither the number of "OR" operators nor the number of "AND" operators correlated significantly with the precision measure. However, the ratio of "OR" operators to "AND" operators does correlate positively. Specifically:

$$\frac{\text{OR}}{\text{AND}} = .29$$

In other words, there is a positive relationship between the number of "OR" operators relative to the number of "AND" operators and precision. The intermediaries appear to begin searches by developing concept classes by linking terms together with "OR" operators. Once these concept classes have been established, they are linked together by "AND' operators for retrieval. It appears that the greater the development of these concept classes (word groups connected by "OR" operators) which are connected by "AND" operators, the higher the precision of the search will be.

In another analysis, it was found that the more display operators the intermediary used, the lower the precision value would be. The correlation between the display operators and the precision was -.45. In this case, it may be that the less sure the intermediary was of their search strategy, the more often the person would take time to display retrieved references. The result being that the display commands would be a measure of

intermediary uncertainty which was reflected in the precision of the search. While studies of this sort are interesting and will be ongoing, they are not central to the understanding of ranking algorithms.

Lost Responses

The evaluated output form from the query processing was usually returned to the project office. The return rate was 69.9%. The 30.1% that were not returned were primarily for the users'personal reasons. However, three queries were lost due to intermediary input errors and computer hardware problems. In both situations, timely return to the user was made impossible.

A software problem caused the loss of forty-five searches from the free representation, but not from the controlled representation. The initial 128 queries were processed normally. Subsequent documents retrieved by the free representation were incorrectly retained and these were not delivered correctly to the users for evaluation. The data were unrecoverable. Thus the data from the free vocabulary tests reflect 128 information request.

MEASURING THE EFFECTIVENESS OF RANKING

The method of evaluation of this study is the coefficient of ranking effectiveness (CRE) and an analysis of the factors affecting the cost of ranking algorithms.

Cooper suggested that the essential function of a retrieval system is not to divide the data base into retrieved versus non-retrieved sets, but rather to establish an ordering among documents based on their relationships to the query. He proposed to measure the effectiveness of a system by its ability to rank order documents - placing relevant documents near the beginning of the list (Cooper, 1968). His measure is the proportional reduction in the number of non-relevant documents that have to be looked at before the query is satisfied, over the number that would have had to be examined if the documents were arranged randomly. Unfortunately, this measure requires knowledge of all the relevant documents in a data base.

An approximation to this measure has been developed which can be computed from a sample without total knowledge of the data base. This measure is not considered to reflect a general assessment of the retrieval system's performance, but rather, it measures specifically the effectiveness of the order in which the output is ranked as opposed to unranked (randomly ordered) output. The Coefficient of Ranking Effectiveness (CRE) is defined as $CRE = \frac{m_r - \bar{R}}{m_r - m_p}$ where m_r is the expected

mean rank of the relevant documents retrieved if the output list is randomly ordered, m_p is the expected mean rank of the relevant documents retrieved if the output list is perfectly ordered, and $\bar{R}$ is the observed mean rank of the relevant documents retrieved. (m_r-m_p) is the distance between the expected mean rank of the relevant documents on a randomly ordered list and their expected mean rank on a perfectly ordered list. The CRE measures the proportion of this distance that is accounted for by the observed mean rank.

If the relevant documents are randomly dispersed throughout the list, then their expected mean rank (over time) will be the same as the mean rank of all the documents. The mean of the number one through n (for n documents on the list) equals $(n+1)/2$.

If there are k relevant documents and they are at the top of the list, then their mean rank would be $(k+1)/2$:

$$CRE = \frac{m_r-\bar{R}}{m_r-m_p}$$

$$= \frac{(n+1)/2-\bar{R}}{(n+1)/2-(k+1)/2}$$

by substitution,

$$= \frac{n+1-2\bar{R}}{n+1-(k+1)}$$

$$= \frac{n+1-2\bar{R}}{n-k}$$

CRE is computed per search. To obtain a mean CRE for a system's performance over a number of searches (for example K searches):

$$\overline{CRE} = \frac{\sum_{i=1}^{k} CRE_i}{K} .$$

CRE ranges from one through negative one, one being perfect ranking, and zero representing random dispersion. A score below zero indicates that the system is performing worse than chance. This means that the relevant documents have a low score, and in order to correct for this, the system would just sort the list low to high instead of high to low.

From its definitional formula it can be seen that CRE is interpretable as percent of error accounted for; it represents the percentage of the total possible improvement (from random to perfect) that has been realized by the system. Since CRE is a linear transform of a mean and $\overline{CRE}$ is a mean of CREs, $\overline{CRE}$ can be expected to be normally distributed as the number of searches on which $\overline{CRE}$ is based increases. [(The expected value of $\overline{CRE}$) = (The expected value of CRE) = 0.]

CRE is relatively insensitive to either the density of relevant documents in the data base or in the retrieval set. Compared to Cooper's measure, it does not require knowledge of all relevant documents in the data base or knowledge of a specific number of documents the user feels will satisfy his query (Cooper, 1968). Another possible measure is Salton's normalized precision (Salton, 1968). Studies show that it is sensitive to the density of relevant documents in the retrieval set and the appearance of non-relevant documents after the location of the last relevant document, and that it is very

sensitive to early occurrences of relevant documents. Hence, it measures something other than is desired in our evaluation, which is aimed at the ability to rank already defined sets. Finally, CRE has an intuitively appealing linearity; a score of 0.5 indicates the mean rank of relevant documents on the list is halfway between what would be expected by chance and what would constitute perfect performance.

Recall measures will not be computed in this study for a number of reasons. One reason is that users do not always want to see all of the relevant documents (Cole and McGill, 1977). Other reasons are (1) the belief that relevance is not an absolute assessment, but rather is a perception of the user, and (2) recall taps aspects of a system's performance which are not within the scope of this study (e.g., coverage of the collection).

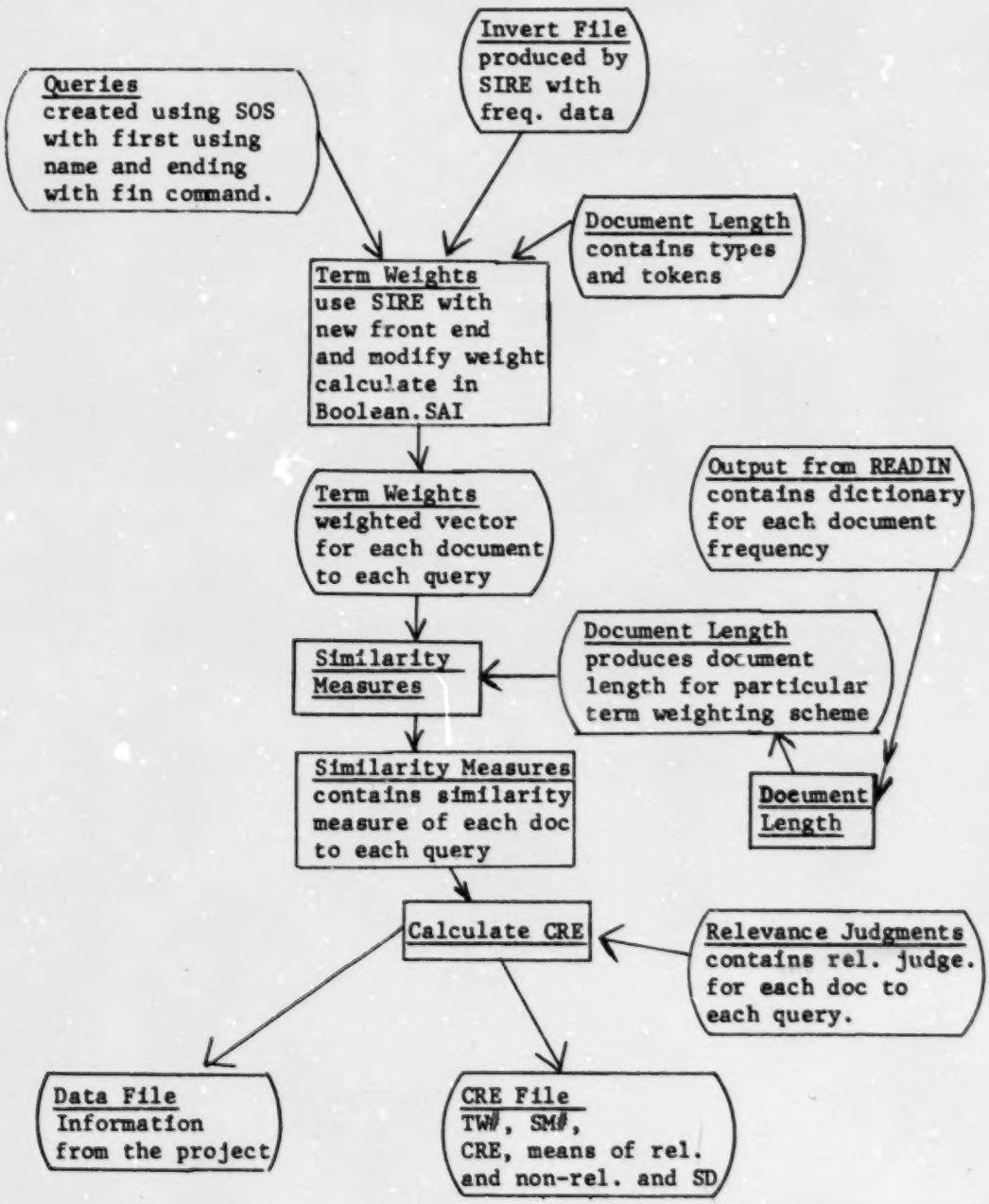
The evaluation required that the relevance data be available and that certain document, collection, and term information be available for analysis. The relevance data was kept in a file organized by search, and then by document retrieval. The ranking algorithms could then be executed and the results compared to the document relevance information. The coefficient of ranking effectiveness and its standard error were calculated directly from this information.

The SIRE system automatically keeps type, token, and sum of square information for terms in documents. That is, for each

document, the information about the number of unique word stems, the number of word stems, the frequency of each stem, and the sum of the square of the frequencies of each term in the document is stored. This is explained in detail in McGill et. al. Document length and collection information were also necessary for this study. This information was captured at the time the document was input to the system and retained for use with the similarity measures. The similarity measures were calculated for each document potentially relevant to a query. The coefficient of ranking effectiveness and associated descriptive information was immediately available and stored for comparative purposes. The procedure is presented in Figure 8 on page 84.

FIGURE 8

CALCULATION OF THE COEFFICIENT
OF RANKING EFFECTIVENESS


Queries
created using SOS with first using name and ending with fin command.
Invert File
produced by SIRE with freq. data
Document Length
contains types and tokens
Term Weights
use SIRE with new front end and modify weight calculate in Boolean.SAI
Term Weights
weighted vector for each document to each query
Output from READIN
contains dictionary for each document frequency
Similarity Measures
Document Length
produces document length for particular term weighting scheme
Similarity Measures
contains similarity measure of each doc to each query
Document Length
Calculate CRE
Relevance Judgments
contains rel. judge. for each doc to each query.
Data File
Information from the project
CRE File
TW#, SM#, CRE, means of rel. and non-rel. and SD

OVERVIEW OF RESULTS

The specific results of this study will be presented in the following two sections. These results indicate little or no differences among term weighting schemes in the controlled vocabulary. This is, of course, an expected outcome. There are significant differences among the similarity measures. There are eighteen measures within two standard errors of the maximum observed value. Thus, while clear differences do exist, there is a class of measures which are not statistically distinguishable. Within the free representation, classes of term weighting and similarity measures are identified which perform significantly better than others. Within classes, no distinction is possible. However, additional cost information is provided to assist in the selection of a ranking algorithm. The cost data are concerned with incremental and relative costs associated with processing and storage. A person using this report is advised to include a consideration of term weighting, similarity measure, and cost.

There are clearly some schemes which are not useful. But of the effective schemes, there is little basis for selecting one scheme over another. Thus, one will be well advised to use simple but effective schemes.

RESULTS USING THE CONTROLLED REPRESENTATION

The $\overline{CRE}$ values for the controlled representation are presented in Table 10. Each cell value is a mean of 68 individual CRE values obtained from the useable queries for this study. See page 87 for Table 10.

Term Weighting Scheme Number

Similarity Measure Number	1	2	6	7	8	9	12	14	22	25	26	27	29	31	Mean
1	.12	.12	.12	.12	.12	.06	.12	.15	.15	.09	.15	.16	.13	.11	.12
2	.07	.13	.07	.07	.13	.12	.13	.14	.14	.14	.16	.18	.14	.17	.13
4	.13	.13	.13	.13	.11	.06	.12	.15	.15	.09	.16	.16	.13	.12	.13
6	.07	.06	.07	.07	.04	.08	.02	.04	.04	.01	.03	.02	.11	.02	.049
7	.11	.13	.11	.11	-.02	.09	.07	.12	.12	.09	.08	.10	.09	.06	.09
11	.12	.13	.12	.12	.11	.12	.14	.17	.16	.16	.15	.18	.13	.18	.14
12	.11	.11	.11	.11	.11	.10	.09	.10	.10	.09	.14	.12	.11	.11	.11
26	.12	-.06	.12	.12	-.01	-.04	-.01	.15	.15	.10	.06	-.02	-.07	.15	.054
27	.12	.07	.12	.12	.06	.12	.05	.07	.08	.05	.12	.12	.11	.07	.091
29	0	-.07	0	0	-.07	-.07	-.07	-.09	-.09	-.10	-.12	-.16	-.08	-.13	-.075
30	.12	.13	.12	.12	.11	.12	.14	.17	.17	.16	.15	.18	.14	.18	.14
31	.11	.10	.11	.11	0	-.01	.01	.08	.08	.08	.07	.09	.09	.13	.075
32	0	.07	0	0	.07	.07	.10	.10	.10	.11	.14	.15	.08	.17	.083
36	.12	.13	.12	.12	.13	.10	.13	.10	.10	.13	.15	.16	.13	.08	.12
40	.12	.13	.12	.12	.10	.12	.15	.13	.13	.15	.16	.16	.12	.13	.13
44	.11	.11	.11	.11	.11	.10	.09	.10	.10	.09	.14	.12	.11	.11	.11
46	.12	.11	.12	.12	.08	.12	.09	.13	.13	.15	.16	.16	.12	.13	.12
50	-.01	.08	-.01	-.01	.07	.03	.07	-.06	-.06	.07	-.02	-.03	-.02	-.05	.0036
52	.12	.12	.12	.12	.12	.12	.14	.16	.16	.14	.17	.18	.13	.19	.14
Random	0	0	0	0	-.01	.02	.02	-.03	.02	.04	.02	-.01	.01	.02	.0071
56	.11	.07	.11	.11	.06	.11	.09	.09	.09	.08	.10	.10	.11	.04	.091
58	.07	0	.07	.07	.03	.06	.02	.08	.08	.06	.07	.07	.07	.04	.056
63	.11	.11	.11	.11	.11	.11	.09	.10	.10	.13	.12	.13	.13	.10	.11
64	.11	.11	.11	.11	.11	.12	.12	.10	.10	.13	.12	.13	.12	.10	.11
Mean	.091	.084	.091	.091	.07	.076	.08	.094	.096	.093	.1	.1	.089	.093	.089

TABLE 10

MATRIX OF $\overline{CRE}$ VALUES

CONTROLLED REPRESENTATION

The analysis of the term weighting schemes is presented in Table 11.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F
AMONG	13	0.027	0.0021	0.55
WITHIN	322	1.2	0.0038	
TOTAL	335	1.3		

TABLE 11

ANALYSIS OF VARIANCE RESULTS

CONTROLLED REPRESENTATION
TERM WEIGHTING SCHEMES

This analysis fails to indicate any significant difference among the term weighting schemes. This is not surprising since the weights for controlled representations are determined by using dichotomous information about the presence or absence of a term. Thus, evaluation of differences is limited to the comparisons of similarity measures.

The values for each cell were examined for differences by similarity measure and for differences by term weighting scheme using one way analysis of variance. The results of the analysis by similarity measures across term weighting schemes are presented in Table 12. A significant difference

at the .01 level is indicated. Thus, similarity measures can be selected which will give significantly better performance on the average than others.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F
AMONG	23	0.89	0.039	33
WITHIN	312	0.36	0.0012	
TOTAL	335	1.3		

TABLE 12

SIMILARITY MEASURE

ANALYSIS OF VARIANCE RESULTS
CONTROLLED REPRESENTATION

In a controlled vocabulary environment, these results indicate that the selection of a term weighting scheme is not an important consideration. The selection of a similarity measure is an important consideration. In order to clarify the selection from among the similarity measures, it is necessary to look for equivalence and disparities in performance.

Tukey's HSD (Honestly Significant Differences) was used to determine significant distinctions between pairs of mean $\overline{CRE}$ values from the similarity measures (KIRK). Table 13 shows the differences among means.

Mean	Frequency of Occurrence	Similarity Measure Number From Table 2	Significantly Different From Means Greater Than	Significantly Different From Means Less Than
-.075	1	29	All Others	None Lower
.0036	1	50	.0516	-.0404
.0071	1	Random	.0551	-.0409
.049	1	6	.097	.001
.054	1	26	.102	.006
.056	1	58	.104	.008
.075	1	31	.123	.027
.083	1	32	.131	.035
.090	1	7	.138	.042
.091	2	27,56	.139	.043
.11	4	12,44,63,64	None Higher	.062
.12	3	1,36,46	None Higher	.072
.13	3	2,4,40	None Higher	.082
.14	3	11,30,52	None Higher	.092

TABLE 13

SIGNIFICANT DIFFERENCES
BETWEEN $\overline{CRE}$ MEANS FOR
SIMILARITY MEASURES
CONTROLLED REPRESENTATION

The critical significant difference interval is calculated at the .05 level with

$$C = q_{.05} \sqrt{\frac{MSw}{n}}$$

where

C = is the critical significant difference limit

MSw = is the mean sequence within cells

n = is the number of term weighting schemes

$q_{.05}$ = is the studentized range statistic estimated at K=24 N-K=∞

By observation the single best result was achieved by the combination of similarity measure 52 with term weighting scheme 31. However, it is clear that similarity measures 11, 30, 52, 2, 4, 40, 1, 36, 46, 12, 44, 63 and 64 may all be performing equivalently. From within this collection there is no reason to expect one to perform better than another. It will be suggested that in the absence of better information, the similarity measure which is the most efficient in storage and computation is currently the most desirable from the above set. Methods of determining storage and processing efficiency will be discussed in a later section.

RESULTS USING THE FREE REPRESENTATION

The $\overline{CRE}$ values for the free representation are presented in Table 14. Each value in the table is the result of 55 individual CRE values obtained from the useable queries.

The analysis of the term weighting schemes is presented in Table 15.

Similarity Measure Number	TERM WEIGHTING SCHEME NUMBER																					
	1	2	6	7	8	9	12	14	18	19	22	24	25	26	27	29	31	32	33	35	37	Mean
1	.059	.117	.120	.209	.105	.221	.030	.110	.088	.223	.107	.207	.167	.030	.069	.068	.098	.056	.186	.093	.127	.119
2	.034	.081	.112	.107	.127	.172	.128	.119	.153	.142	.094	.147	.059	.127	.127	.065	.124	.112	.175	.125	.158	.118
4	.060	.053	.197	.209	.106	.221	.037	.110	.087	.221	.108	.207	.167	.125	.063	.068	.098	.056	.186	.093	.128	.124
6	.035	₹011	₹030	₹027	₹025	₹034	.068	.065	.062	₹020	₹037	.025	₹025	.060	.032	.085	.084	.058	₹023	₹019	.068	.019
7	.059	.113	.135	.119	.131	.159	.034	.062	.063	.147	.059	.135	.119	.036	.068	.124	.070	.116	.138	.125	.091	.100
11	.042	.086	.151	.137	.127	.172	.128	.123	.141	.156	.108	.164	.078	.032	.127	.061	.121	.112	.175	.125	.157	.120
12	.081	.147	.182	.164	0	.173	.101	.093	.101	.171	.092	.161	.149	.072	.091	.081	.105	.013	.161	0	.119	.107
26	.055	.102	.175	.156	.137	.189	.034	.103	.035	.182	.109	.169	.153	.036	.038	₹055	.128	.072	.034	.123	.149	.101
27	.071	.091	.074	.073	.096	.088	.067	.055	.073	.069	.068	.050	.050	.042	.100	.083	.058	.048	.111	.092	.058	.072
29	0	₹073	₹091	₹091	₹144	₹165	₹126	₹121	₹136	₹132	₹055	₹109	₹052	₹118	₹110	₹064	₹105	₹075	₹159	₹122	₹139	₹104
30	.042	.116	.152	.137	.127	.173	.125	.119	.152	.154	.093	.158	.076	.029	.128	.061	.120	.112	.175	.125	.158	.121
31	.059	₹049	.127	.139	.118	.184	₹016	.071	.020	.114	.039	.113	.069	0	.038	.048	₹016	.114	.165	.111	.011	.069
32	0	.026	.091	.089	.143	.162	.085	.108	.137	.127	.055	.122	.035	.002	.109	.065	.105	.105	.164	.123	.161	.096
36	.061	.091	.188	.187	.053	.190	.101	.109	.132	.208	.058	.180	.133	.128	.116	.061	.073	.096	.177	.032	.089	.117
40	.060	.082	.181	.169	.114	.166	.086	.092	.123	.179	.082	.169	.147	.029	.068	.069	.129	.093	.147	.101	.148	.116
44	.081	.150	.182	.164	.003	.173	.101	.093	.101	.171	.092	.161	.149	.072	.091	.081	.105	.013	.161	0	.119	.108
46	.061	.105	.175	.168	.021	.173	.101	.090	.116	.176	.081	.168	.147	.042	.072	.101	.130	.038	.139	.010	.149	.108
50	₹064	.047	₹124	₹118	₹141	₹088	.008	₹063	.009	₹143	₹038	₹107	₹108	.033	.018	.030	₹053	₹077	₹003	₹131	₹074	₹057
52	.042	.034	.119	.120	.102	.115	.057	.124	.114	.132	.078	.142	.082	.034	.100	.045	.086	.050	.133	.086	.111	.091
Random	.049	₹013	₹006	₹050	.039	.018	.021	₹034	₹078	₹031	.005	.036	.034	.010	₹071	₹011	₹043	₹021	₹067	₹085	₹049	₹017
56	.059	.091	.107	.101	.087	.099	.060	.060	.065	.097	.029	.089	.095	.062	.069	.081	.035	.061	.097	.084	.053	.075
58	.059	.117	.111	.105	.118	.104	.034	.060	.045	.106	.030	.094	.093	.029	.072	.090	.033	.028	.103	.076	.050	.074
63	.047	.059	.042	.042	.037	.058	.063	.057	.058	.044	.048	.056	.055	.059	.057	.060	.062	.067	.058	.039	.052	.053
64	.047	.059	.045	.044	.039	.058	.063	.052	.058	.047	.048	.057	.056	.059	.057	.060	.057	.066	.060	.039	.052	.053
Mean	.046	.068	.101	.098	.063	.116	.058	.069	.072	.106	.056	.108	.080	.043	.064	.057	.067	.055	.104	.052	.081	.0744

TABLE 14

MATRIX OF $\overline{CRE}$ VALUES

FREE REPRESENTATION

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F
AMONG	20	.24096	.0120	2.353
WITHIN	483	2.471	.0051	
TOTAL	503	2.714		

TABLE 15

ANALYSIS OF VARIANCE RESULTS

TERM WEIGHTING SCHEMES

FREE REPRESENTATION

The analysis indicates a significant difference among the term weighting schemes at the .01 level. Again using Tukey's Honestly Significant Difference, the individual means were examined to determine if significant distinctions can be made between pairs of mean $\overline{CRE}$ values.

The significant difference value is

$$C = q_{.05}\sqrt{\frac{MSw}{N}}$$

with $q_{.05}$ at $K = 21$, $N-K = 483$ $= 5.05\sqrt{\frac{.0051}{21}} = .07869$

The mean $\overline{CRE}$ values range from .043 to .116. Thus, no significant difference is found between individual pairs of term weighting schemes. This situation can arise when linear combinations of term weighting schemes are significantly

different but the individual pairs of term weighting schemes are not significantly different.

The analysis of the similarity measures is presented in Table 16.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F
BETWEEN	23	1.698	.0738	34.9133
WITHIN	480	1.015	.0021	
TOTAL	503	2.713		

TABLE 16

ANALYSIS OF VARIANCE RESULTS

SIMILARITY MEASURES

FREE REPRESENTATIONS

The analysis indicates a significant difference among the similarity measures at the .01 level. Tukey's Honestly Significant Difference was used to examine the significant differences. The significant difference value is

$$C = q_{.05} \sqrt{\frac{MSw}{N}}$$

with $q_{.05}$ at $K = 24$, $N-K = 480 = 5.17$

$$C = 5.17 \sqrt{.0021/24} = .04836$$

Table 17 shows the similarity measures in order along with the definitions of those measures which are within the Honestly Significant Difference. Since the number of similarity measures which do not indicate an honest difference is large (14) then there is justification for selecting a similarity measure from among these top rated measures based on its ease of calculation and quantity of storage required.

EFFICIENCY CONSIDERATIONS

In order to decide on which ranking algorithm(s) to implement, one would like an indication of the cost to implement and execute the algorithm in addition to its effectiveness. A model which analyses costs is dependent on the specifics of the computer installation on which the algorithm is to be implemented. Differences in hardware and/or software can make significant changes in costs.

The characteristics identified here provide a weak ordering of the ranking algorithms. Factors which affect costs are identified and each of the components of the algorithms is placed in this framework. The characteristics then indicate the relative cost of the algorithm.

A major component of ranking algorithms is the cost associated with the TW, although the SM also affects the cost. The following will describe the considerations which appear to

Mean	Frequency of Occurrence	Similarity Measure Number from Table 2	Significantly Different from Means GREATER than	Significantly Different from Means LESS than
-.104	1	29	All Other Means	None Lower
-.057	1	50	-.009	None Lower
-.017	1	Random	.031	-.065
.019	1	6	.067	-.029
.053	2	63,64	.101	.005
.069	1	31	.117	.021
.072	1	27	.12	.024
.074	1	58	.122	.026
.075	1	56	.123	.027
.091	1	52	None Higher	.043
.096	1	32	None Higher	.048
.1	1	7	None Higher	.052
.101	1	26	None Higher	.053
.107	1	12	None Higher	.059
.108	2	44,46	None Higher	.060
.116	1	40	None Higher	.068
.117	1	36	None Higher	.069
.118	1	2	None Higher	.070
.119	1	1	None Higher	.071
.12	1	11	None Higher	.072
.121	1	30	None Higher	.073
.124	1	4	None Higher	.076

TABLE 17

SIGNIFICANT DIFFERENCES BETWEEN CRE MEANS FOR SIMILARITY MEASURES

FREE REPRESENTATION

be the most important in determining the cost of a ranking algorithm. Two considerations are paramount: (1) the processing requirements of the algorithm and (2) the storage costs of the algorithm. An assumption of this analysis is that the Term Weighting for each term is calculated and stored only once.

COSTS OF TWs

The processing costs for the term weighting schemes are largely determined by whether the specific weighting of the TW requires one or two passes through the data base. Many of the weighting schemes required two passes. For example f_{in}/F_i required one pass to calculate F_i (a collection statistic) for each term and a second pass to determine the weight for each term. On the other hand, log $(f_{in}+1)$ can be calculated on the first pass because no collection information is required. Of course, it is possible to trade processing time for storage and do the operation f_{in}/F_i at retrieval. However, this was not examined in this study.

The incremental storage costs are determined by the number of additional storage locations needed to store the actual weights. The two choices are: (1) a weight for each unique term in the dictionary, or (2) a weight for each unique term in the data base. The former required weights for each unique term in the dictionary (generally a collection statistic) for the free representation, 10,000 storage units, and the weights for each unique term in the data base required 170,000 storage units.

The null TW requires no additional storage or processing. That is, the unweighted scheme represents the lower limit of cost of term weighting schemes. Costs of term weighting schemes are presented by class with those costing the most appearing first. Within classes there are variations due to the complexity of calculation. These are not considered critical, since specific calculations are minor in comparison to calculations conducted on a data base.

CATEGORY 1 - Two Passes of Data Base Required.
One Storage Unit per Unique Term
One Storage Unit per Document

$$f_{in}/F_i \;,\quad f_{in}\frac{F_i}{d_i} \;,\quad \frac{t_n\, d_i}{F_i - f_{in}} \;,\quad f_{in}.\log\frac{(K)}{F_i} \;,\quad \frac{f_{in}}{\log F_i}$$

$$\frac{1}{T_n\, d_i} \;,\quad \frac{1}{\log\,(t_n.\, d_i)} \;,\quad \frac{1}{t_n} - \frac{d_i}{D} \;,\quad \frac{\frac{1}{t_n} - \frac{d_i}{D}}{\sqrt{d_i/D}} \;,\quad \frac{f_{in}}{k_n\, F_i} \;,$$

$$\frac{f_{in}}{\log\,(k_n\, F_i)} \;,\quad \frac{f_{in}}{k_n} - \frac{F_i}{K} \qquad \frac{\frac{f_{in}}{k_n} - \frac{F_i}{K}}{\sqrt{\frac{F_i}{K}}}$$

CATEGORY 2 - One Pass
One Storage Unique/Unique Term - Document

$$\frac{1}{t_n}, \quad f_{in}, \log f_{in}, \quad \frac{f_{in}}{k_n}, \quad \frac{f_{in}}{\log k_n}$$

CATEGORY 3 - One Pass
One Storage Unit/Unique Term - Dictionary

$$\frac{1}{d_i}, \quad \log\left(\frac{N}{d_i}\right)$$

COST OF SMs

A major factor in the costs of a similarity measure is the need by the measure of summary statistics of the document, such as Σx_i. If this is required, then an additional storage unit is required for each document in the data base.

The use of summary information such as document length by the SM may have an interaction affect when combined with certain TWs. In particular, the use of either $1/d_i$ or $\log (N/d_i)$ with such a similarity measure would alter these TWs from one-pass to two-pass TWs. The second pass through the data base is required to calculate the length of the document for the weighting scheme.

The discussion of costs of ranking algorithms indicates the observed factors which affect cost. Exact cost data is

installation dependent and thus not generally useful. The established categories provide a weak ordering of the different ranking algorithms in terms of processing costs and storage requirements. The reader should pay particular attention to the interaction of term weighting and similarity measures in determining the cost of any particular ranking algorithm.

CONCLUSIONS

This study indicates that many of the ranking algorithms currently in use or suggested as effective methods for ordering output are, in fact, equivalent. Further, as one would expect, the term weighting schemes in the controlled environment are simply not important. This is evident from the lack of significance shown by the analysis of variance given that the unweighted weighting scheme is isolated.

Term weighting in the free text environment is significant. However, the use of Tukey's Honestly Significant Difference fails to indicate a significant difference between pairs of the term weighting schemes.

Similarity measures in both the free and the controlled environment are significantly different. Classes of measures which were found to be equivalent have been presented. These top rated measures are still disappointing. That is, by observation the best ranking scheme in the controlled environment had a $\overline{CRE}$ of .19 and in the free environment the top scheme had a $\overline{CRE}$ of .22. In other words, in both instances the ranking algorithm was able to improve the order in which documents appeared by about 20%. Thus, 80% of the potential benefits from a ranking algorithm is not yet realized. These results do agree in general with those attained by Noreault who found a 35% improvement. The data seem to indicate that the methods of ranking are not using variables which allow

truly effective ranking (at least in a Boolean Environment). This may unfortunately be a major factor in one's ability to create truly effective systems. Maron (1979) has recently stated that:

> "Two valued thinking about indexing (and retrieval) leads system designers to worry about thresholds, cutoff values, and depth of indexing in order to insure that the two-valued decisions are optimal for the patrons for whom the system is designed to serve. But these days with the growth of very large files and especially with the growth of on-line, interactive document retrieval systems, perhaps it is most rational to build systems that provide maximum flexibility for each patron. This means that designers should build systems which rank the documents, relative to an input query, by probability (or degree) of satisfaction, and set no preestablished cutoff thresholds. Instead of binary indexing, we recommend the use of weighted indexing and ranking the output documents."

His stated goal is ranking the output, but unfortunately it is not clear from this study that the use of weighted indexing will provide his desired results. Ongoing analyses of this ranking data may help to indicate the variables or variable types that contribute positively to the ranking process.

The results of this study have raised many questions. At a very basic level, one needs to understand the data generated by the study of the overlap among retrieved sets of documents. The observed data indicate that the specific documents retrieved by an individual representation are different from those retrieved by another representation

even though the original information need is identical. Further, the data show that the overlap in documents retrieved by different searchers is small. That is, in response to the same information need, different searchers appear to be retrieving different documents. This may be an artifact of this particular study, or it may be a general situation. It would seem that only additional data will answer this.

In either case, there is a sense of uneasiness associated with conducting a study which examines specific documents when there is question about the factors underlying the selection of the documents. Fortunately, this did not detract from the methodology used in this study. The effectiveness of the ranking algorithms measures changes in order after the set is retrieved.

The data also suggests that the use of frequency information, whether by document or by collection, is limited in its current ability to rank documents. The term weighting schemes and similarity measures were selected to represent the schemes available in the open literature. Thus, a significant increase in the ability of a ranking algorithm is not likely to occur by a calculation which employs some rearrangement of these frequency variables. Rather, it seems that new factors will have to be identified to resolve a significant portion of the 80% benefit not attained by current algorithms.

The current study does show that documents can be rearranged to aid the user. The rearrangement will, in general, move relevant documents toward the beginning of the list of output documents. The overall effect is beneficial to the user. Further, if the algorithm is selected using the data about effectiveness and efficiency, then the cost to the system can be minimized while giving all the benefit we know how to provide at this time.

REFERENCES

ARTANDI, S.; WOLF, E.H. 1969. "The Effectiveness of Automatically Generated Weights and Links". American Documentation, 20(3):192-202 (1969).

BALL, G.H. 1965. "Data Analysis in the Social Sciences: What About the Details?". Proceedings of AFIPS Fall Joint Computer Conference, 1965, 533-559.

BOOKSTEIN, A. 1977. Personal Communication, 1977.

BOOKSTEIN, A.; COOPER, W.S. 1976. "A General Mathematical Model for Information Retrieval Systems". Library Quarterly, 1976 April; 46(2):153-167.

CAGAN, C. 1970. "A Highly Associative Document Retrieval System". Journal of the American Society for Information Science. 21:330-337 (1970).

CARROLL, J.M.; ROELOFFS, R. 1969. "Computer Selection of Keywords Using Word-Frequency Analysis". American Documentation, 20(3):227-233 (1969).

CLEVELAND, D.B. 1976. "An n-Dimensional Retrieval Model". Journal of the American Society for Information Science. 27(5/6):342-347 (1976).

CLEVERDON, C.; KEEN, M. 1966. "Factors Determining the Performance of Indexing Systems", Volume 2. Test Results. ASLIB, Cranfield Research Project, 1966.

COLE, E.; McGILL, M.J. 1977. "Professional Activities of Research Scientists Aided by CANISDI; An Approach to the Information Client". Syracuse, NY: School of Information Studies, Syracuse University; 1977.

COOPER, W.S. 1968. "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering of Retrieval Systems". American Documentation. 1968 January; 19(1):30-41.

COOPER, W.S. 1970. "On Deriving Design Equations for Information Retrieval Systems". Journal of the American Society for Information Science. 21(6):385-395 (1970).

CORMACK, R.M. 1971. "A Review of Classification". Journal of the Royal Statistical Association, Series A. 134:321-353 (1971).

EDMUNDSON, H.P.; WYLLYS, R.E. 1961. "Automatic Abstracting and Indexing - Survey and Recommendations". Communications of the ACM, 4(5):226-234 (1961).

GEBHARDT, F. 1975. "A Simple Probabilistic Model for Relevance Assessment of Documents". Information Processing and Management. 1975;11(1/1):59-65.

HARTER, S. 1975. "A Probabilistic Approach to Automatic Keyword Indexing: Part I". Journal of the American Society for Information Science. 26(4):197-206 (1975).

HARTER, S. 1975. "A Probabilistic Approach to Automatic Keyword Indexing: Part II". Journal of the American Society for Information Science. 26(5):280-289 (1975).

KATTER, R. 1967. "A Study of Document Representation: Multidimensional Scaling of Index Terms". SDC-Final Report, August 31, 1967.

KATZER, J. 1971. Syracuse University Psychological Abstracts Retrieval Service. Final Report. "Large Scale Information Processing Systems, Section V: Cost-Benefit Analysis". Syracuse University, School of Library Science, 1971.

KEEN, E.M. 1973. "The Aberystwyth Index Languages Test". Journal of Documentation. 29(1):1-35 (1973).

KIRK, ROGER E. 1968. "Experimental Design." Brooks/Cole Publishers Inc., New York (1968).

KUHNS, J.L. 1965. "The Continuum of Coefficients of Association". In: Stevens, M.E., Giuliano, V.E. and Heilprin, L.B. (eds) Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Washington, 1974 (NBS Misc. Publ. Mo. 269, 1965), 33-40.

LANCASTER, F.W.; FAYEN, E.G. 1973. "Information Retrieval On-line". Los Angeles, CA: Melville Publishing Co.; 1973. 414.

LERMAN, I.C. 1970. "Les Bases de la Classification Automatique." Paris: Gauthier-Villars, 1970.

LUHN, H.P. 1957. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information". IBM Journal of Research and Development, 1(4):309-317 (1957).

MARON, M.E. 1979. "Depth of Indexing". Journal of the American Society for Information Science. Volume 30, Number 4, July 1979, p.227.

MARON, M.E.; KUHNS, J.L. 1960. "On Relevance Probabilistic Indexing and Information Retrieval". Journal of the ACM. 7(3):216-244 (1960).

McCARN, D. 1976. Personal Communication.

McGILL, M.J.; SMITH, L.C.; DAVIDSON, S.; NOREAULT, T. 1976. "Syracuse Information Retrieval Experiment (SIRE): Design of an On-line Bibliographic Retrieval System". SIGIR FORUM of the ACM, X(4):37-44 (1976).

MINKER, J.; WILSON, G.A.; ZIMMERMAN, B.H. 1972. "An Evaluation of Query Expansion by the Addition of Clustered Terms for a Document Retrieval System". Information Storage and Retrieval. 8:329-348, (1972).

NOREAULT, T.; KOLL, M.; McGILL, M.J. "Automatic Ranked Output from Boolean Searches in SIRE". JASIS 1977, 28:333-339.

ORLOCCI, L. 1969. "Information Theory Models for Hierarchic and Non-Hierarchic Classifications". In: Cole, A.J. (ed) Numerical Taxonomy, Proceedings of Colloquium in Numerical Taxonomy at University of St. Andrews, Scotland, 1968. NY: Academic Press, 1969, 148-164.

OVERALL, J.E.; KLETT, C.J. 1972. "Applied Multivariate Analysis" Englewood Cliffs: Prentice Hall, 1972.

REITSMA, K.; SAGALYN, J. 1968. "Correlation Measures". In: Information Storage and Retrieval. ISR Report No. 13, 1968.

RICKMAN, J.T. 1972. "Automatic Storage and Retrieval for On-Line Abstract Collections". Doctoral Dissertation, Pullman WA: Department of Computer Science, Washington, State University; 1972.

ROBERTSON, S.E. 1974. "Specificity and Weighted Retrieval". Journal of Documentation. 30(1):41-46 (1974).

SAGER, W.K.H.; LOCKEMANN, P.C. 1976. "Classification of Ranking Algorithms". International Forum on Information and Documentation. 1(4):2-25, (1976).

SALTON, G.; LESK, M.E. 1968. "Computer Evaluation of Indexing and Text Processing". Journal of the ACM. 15(5):8-36 (1968).

SALTON, G. 1968. "Automatic Information Organization and Retrieval". New York: McGraw-Hill Book Company 1968.

SALTON, G. 1969. "A Comparison Between Manual and Automatic Indexing Methods". American Documentation. 20(1):61-71 (1969).

SALTON, G.; YANG, C.S. 1973. "On the Specification of Term Values in Automatic Indexing". Journal of Documentation. 29(4):351-371, 1973.

SALTON, G. 1975. "Dynamic Information and Library Processing". Prentice-Hall Inc. Englewood Cliffs, NJ. pp. 504.

SALTON, G.; YANG, C.S., YU, C.T. 1975. "A Theory of Term Importance in Automatic Text Analysis". Journal of the American Society for Information Science, 26(1):33-44 (1975).

SALTON, G.; WONG, A.; YU, C.T. 1976. "Automatic Indexing Using Term Discrimination and Term Precision Measurements". Information Processing and Management. 12:43-51 (1976).

SARACEVIC, T. 1968. "An Inquiry into Testing of Information Retrieval Systems". Comparative Systems Laboratory Technical Reports No. CSL:TR-FINAL 1 to 3. Cleveland: Center for Documentation and Communication Research, Case Western Reserve University, 1968.

SNEATH, P.H.A.; SOKAL, R.R. 1973. "Numerical Taxonomy: The Principles and Practice of Numerical Classification". San Francisco, CA. W.H. Freeman and Co., 1973).

SPARCK JONES, K.; JACKSON, D.M. 1970. "The Use of Automatically-Obtained Keyword Classifications for Information Retrieval". Information Storage and Retrieval. 5(4):175-202 (1970).

SPARCK JONES, K. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation, 28:11-21 (1972).

SPARCK JONES, K. 1973. "Index Term Weighting". Information Storage and Retrieval. 9:619-633 (1973).

SPARCK JONES, K. 1974. "Automatic Indexing: A State-of-the-Art Review". Computer Laboratory, University of Cambridge, 1974.

SVENONIOUS, E. 1972. "An Experiment in Index Term Frequency". Journal of the American Society for Information Science. 23:109-121, (1972).

SWETS, J.A. 1967. "Effectiveness of Information Retrieval Methods", Bolt Beranek and Newman Rept. 1499. Cambridge, Mass. April 1967.

SWITZER, P. 1964. "Vector Images in Information Retrieval". In Statistical Association Methods for Mechanical Documentation, Symposium Proceedings, Washington, D.C. 1964. (NBS Misc. Publ. 269, 1965). pp. 163-171.

TAGUE, J. "An Evaluation of Statistical Association Measures". Proceedings of American Documentation, 1966, 391-397.

TARS, A. 1976. "Stemming As A System Design Consideration". ACM SIGIR FORUM, Spring 1976, pp. 9-16.

TORGERSON, W.S. 1958. "Theory and Methods of Scaling". NY: John Wiley and Sons, Inc. 1958.

van RIJSBERGEN, C.J. 1975. "Information Retrieval". London: Butterworths, 1975.

van RIJSBERGEN, C.J. 1977. "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval". Journal of Documentation, 33(2):106-119, (1977).

YU, C.T.; SALTON, G. 1977. "Effective Information Retrieval Using Term Accuracy". Communications of the ACM, 20(3):135-142, (1977).

APPENDIX A

NSF RANKING PROJECT

Instructions to Searchers

1. Make sure first digit of each Information Requirement Statement (IRS) I.D. number is yours.

2. Process the IRSs in the order given to you.

3. a. IRSs marked ERICF (having ID numbers ending in 0) are to be searched against the ERICF file. This is a dictionary of stems from document titles and abstracts.

   b. IRSs marked ERICC (having ID numbers ending in 1) are to be searched against the ERICC file. This is an indexer-assigned, controlled vocabulary.

4. Listings of both dictionaries are available in hard copy.

5. a. Search as you would under normal working conditions.
   b. Try to provide the user with high recall. That is, lean towards inclusion, rather than exclusion, of possibly relevant documents.
   c. You may occasionally process the same request twice, once against ERICF and once against ERICC. Treat these as independent requests. Treat the second IRS as if you had not read it before.

6. Use the IRS forms as worksheets.

7. Refer to SIRE Instruction Sheet for aid using SIRE.

8. When satisfied (or at least done) with the set of documents retrieved for an IRS, issue the "DONE" command with the full IRS ID number.

9. When done, rip off paper from your terminal and insert in folder.

10. When a terminal session is through, return IRS forms for completed searches, along with all terminal output in file folders.

NSF RANKING PROJECT

SIRE Instruction Sheet

Boolean query

Submits a Boolean logic query. An "and", "or" or "not" must appear between each search term. The Boolean operators are processed left to right.

Save N

Saves the results of previous search in location N. N is an integer value between 1 and 5 inclusive. A saved set may be used as a term in a later search by using *N in the query.

List

Lists the document numbers retrieved by the previous search.

TA

"Type Abstract" - Types complete bibliographic citation plus the abstract and descriptors for the N+h ranked document. N may also be a range of ranked documents, e.g. "TA 1-5" types the first five documents from the retrieved set.

TS N

"Type Short" - Same as TA except abstract and descriptors are not printed.

DONE Query Number

When finishing a search, issue this command before next search.

END

Ends SIRE execution.

Switch file name

ERICF - for title and abstract

ERICC - for controlled vocabulary

In the controlled vocabulary, the words in a single search term are separated using a "/". If the word is not in the dictionary, it will be echoes with 0 postings. If it is in the dictionary, the code for the word will be echoes.

APPENDIX B

NSF RANKING PROJECT

Instructions for Tracking Queries

1. Fill in the following chart as appropriate:

	N|NNNN|N	Date IRS Recvd	Date Out To Searcher	Date In From Searcher	Date Out To User	Date In From User
1.	0001 0					
2.	0001 1					
3.	0002 0					
4.	0002 1					
.	.					
.	.					
.	.					

2. When Information Requirement Forms (IRS) are received from user, assign each an accession number (NNNN).
3. Make 2 copies of each IRS.
4. File master by accession number.
5. Label one copy "ERICF" and add "0" to end of accession number. Label other copy "ERICC" and add "1" to end of accession number.
6. For each copy, individually, assign a random number to the front of the accession number. Roll die:

IF	ASSIGN
1, 2	1
3, 4	2
5, 6	3

7. Create a file folder for each copy, labelled N\|NNNN\|N Store separately.
8. Give folder to appropriate searcher.
9. When searcher returns folder, prepare printed output for return to user. (Cut, burst, staple and assemble with original IRS).
10. When relevance assesments are returned from user, add to the NNNN 0 folder.

SYRACUSE UNIVERSITY
SCHOOL OF INFORMATION STUDIES

111 EUCLID AVENUE · SYRACUSE, NEW YORK 13210 · PHONE (315) 423-2911

NSF INFORMATION RETRIEVAL PROJECT

We will conduct a computer search of four Computer Index Journals in Education data bases for you if you will simply tell us what it is you would like us to search for and tell us how we did after the search. You will have access to the data bases created by the clearinghouses on EDUCATIONAL MANAGEMENT, TEACHER EDUCATION, TESTS, MEASUREMENT AND EVALUATION and INFORMATION RESOURCES. These data bases run from 1975 through items made available less than a month ago.

The attached form is for you to describe the topic of interest. Don't worry about trying to say it in computerese. You say it in English. We have trained people to make sure that your search is conducted professionally. In about a week you will receive a list of references and abstracts found on your topic. You will then be asked to indicate which of these are pertinent to your interests. That is all there is to it. You keep one copy of the computer output, and return one copy to us telling us which references are pertinent.

Name: _____

Address: _____

_____

_____

Instructions to Participants

Choose a specific or general topic, in Education, you need information on right now. If you are doing a paper or planning a talk you probably have a topic in mind. If you don't have any topic you are working on, consider one you are familiar with.

In order to acquire this information for your topic we want you to write down your information requirements as if you were talking to a colleague who understood the field as well as you do.

Make this statement as precise and concise as possible. This statement should be clear enough so that any person with a knowledge of education would, on the basis of this statement alone, be able to pick out sources which would be of interest to you.

In 2 - 4 sentences describe the information you want:

_____

_____

_____

_____

_____

_____

In a short time you will receive a list of references and abstracts that have been retrieved by computer from a data base consisting of document references from Computer Index Journals in Education. You will be asked at that time to let us know which of these references you think would be pertinent to your interest. Thank you for your cooperation.

NSF Information Retrieval Project
School of Information Studies
113 Euclid Avenue
Syracuse, New York 13210
(315) 423-4522

NSF INFORMATION RETRIEVAL PROJECT

INSTRUCTIONS TO PARTICIPANTS

Attached you will find a copy of your interest statement and two copies of a list of references. Each reference consists of seven parts:

DN - Document identification number

TI - Title

AU - Author

SO - The source of the reference (eg; The title of the journal in which the article appeared)

AB - Abstract

DT - Date

DE - Descriptors of the reference

List (a) is to be used as part of the study and should be returned. Copy (b) is yours to keep.

From each citation and abstract you form an idea of what that particular document (book, article, report) is about. Compare this to your interest statement, and for each document listed, decide how closely that document is related to your topic. Based on the information in front of you is the document relevant to your topic or not relevant to what you had in mind.

Judge on a scale from 1 to 4:

1 - Definitely relevant to your topic.

2 - probably relevant to your topic.

3 - Probably not relevant to your topic.

4 - Definitely not relevant to your topic.

Place this number in the box next to each reference.

SCHOOL OF INFORMATION STUDIES * SYRACUSE UNIVERSITY

Memorandum

To

Date

Subject COMPUTER SEARCH

Recently the NSF Information Retrieval Project prepared a computer search for you.

Part B was for your reference and Part A was to have a relevance judgment and be returned to us. As yet we have not received Part A from you.

It would be appreciated if you would complete the judgement and return Part A to

NSF Information Retrieval Project
School of Information Studies
113 Euclid Avenue
Syracuse, New York 13210
(315) 423-4522

Thank you for your prompt attention to this request.

Michael J. McGill